Understanding Ethnolinguistic Differences: The Roles of Geography and Trade

Andrew Dickens[†]

June 2021

Abstract

I study the role of trade on inter-ethnic linguistic differences in the long run. I hypothesize that the geographic environment of neighbouring ethnic groups determines their potential gains from trade, and that the frequency of inter-ethnic trade—and resulting social interactions—shape the co-evolution of language. As a test of this hypothesis, I build a georeferenced dataset to examine the border region of spatially adjacent ethnic groups, together with variation in the set of potentially cultivatable crops at the onset of the Columbian Exchange, to identify how variation in land productivity impacts linguistic differences between adjacent ethnic groups. I find that ethnic groups separated across geographic regions with high variation in land productivity are more similar in language than groups separated across more homogeneous regions. I develop a model to theoretically ground this link between land productivity variation and inter-ethnic trade, and provide empirical evidence in support of this mechanism, including direct evidence of a causal link between land productivity variation and an ethnic group's reliance on trade for food and subsistence in pre-modern times.

Keywords: Ethnolinguistic diversity, Linguistic differences, Culture, Social interactions, Trade, Geography

JEL Classification Codes: N50, O10, Z13, J15, Z10

[†] Brock University, Department of Economics, St. Catharines, ON. E-mail: adickens@brocku.ca. A special thanks to Nippe Lagerlöf for his detailed feedback and support. I am very grateful for comments from the editor, Hans-Joachim Voth, and two anonymous referees. This research has also benefited from discussion with Tasso Adamopoulos, Chris Bidner, Arthur Blouin, Miguel Cardoso, Greg Casey, James Fenske, Marc Klemp, Oded Galor, Dietrich Vollrath, and seminar participants at Brock University, Wilfrid Laurier University, York University, CEA Annual Meeting (McGill), DEGIT XXIV (Southern Denmark) and ASREC (Lund). I also thank Stelios Michalopoulos for generously sharing data with me.

1 Introduction

It is well understood that history shapes the evolution of culture (Nunn, 2012). The long arc of history then puts great importance on understanding the channels through which culture persists and changes, given the mounting evidence that history and culture impact comparative economic development today (Spolaore and Wacziarg, 2009; Comin et al., 2010; Putterman and Weil, 2010; Chanda et al., 2014). While there is a large literature linking major historical events to episodes of cultural change, there is little empirical evidence on the economic mechanisms that shape culture.[1] The aim of this research is to shed light on an unexplored channel of cultural change: inter-ethnic trade.

To this end, I propose a two-part hypothesis: the geographic environment of neighbouring ethnic groups determines their potential gains from trade, and the frequency of inter-ethnic trade mediates the process of cultural persistence and change. To illustrate the logic of the first part, I begin with the observation that land quality determines a society's productive capabilities in a pre-industrial phase of development. The potential gains from trade are highest in regions with high variation in land productivity, since high variation regions give life to a wide range of producible goods and an opportunity for specialization. I provide a Heckscher-Ohlin-style model to make this point, where the gains from trade between two ethnic groups are increasing in between-group variation in agricultural land endowments. This simple model builds on a long-standing observation that the "diversity of the ecosystem thus promotes diversity in production and, with it, exchange over space" (Bates, 2010, p. 21).

The second part of the hypothesis—the mediating effects of trade on culture—relates to the fact that, in pre-modern times, trade fostered communication networks and served as a social tie between spatially and culturally distinct ethnic groups (Koetsier, 2019). The theoretical link between social mechanisms of this type and the evolution of culture is well documented in the literature (Centola et al., 2007). Repeated social interactions build inter-ethnic trust and tolerance, the result of which can be self-stabilizing when the complementarities of inter-ethnic trade are costly to replicate otherwise (Jha, 2013, 2018).

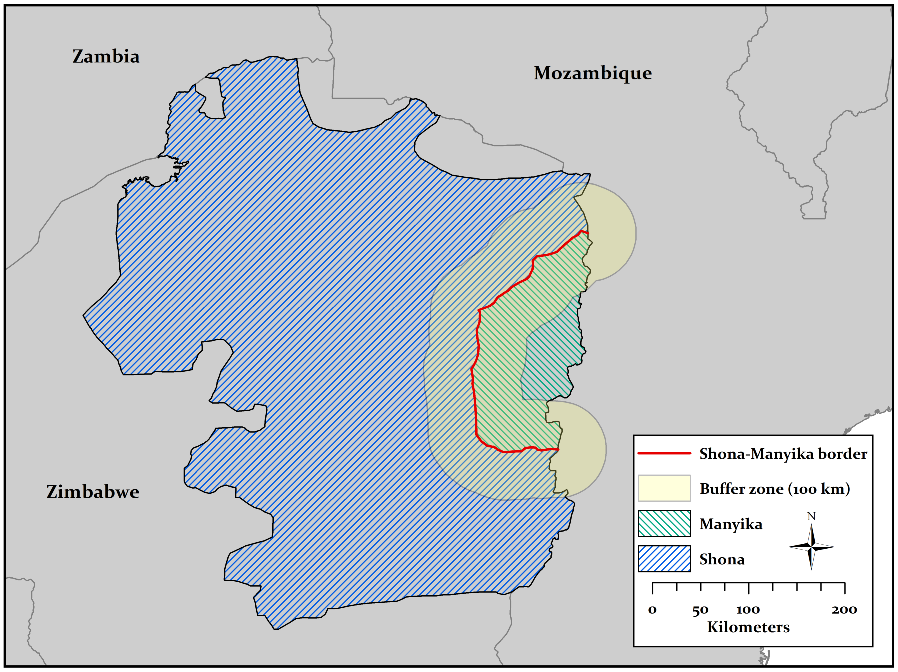
To test this hypothesis, I use a variety of data and methods. I start with the *Ethnologue* and construct a georeferenced dataset of spatially adjacent ethnic groups from across the world.[2] I extract the border segment connecting each adjacent pair, and construct a buffer zone around each border segment as my unit of observation. Each buffer zone is used as a topographic lens to observe the geographic region that links a spatially adjacent group pair (e.g., see Figure 1).

For the independent variable of interest, I measure variation in land productivity within

[1] See Voigtlander and Voth (2012), Giuliano and Nunn (2013), Alesina et al. (2013), Becker et al. (2016) and Guiso et al. (2016) for evidence of cultural persistence and change due to factors rooted deep in history.

[2] An ethnic group is a social grouping of people that is rooted in the belief of shared ancestry. A basic feature of an ethnic group is some form of a cultural community, often manifested in a common language, where a sense of solidarity within the community unites its members (Fearon, 2003).

Figure 1: Unit of Observation: Border Buffer Zone



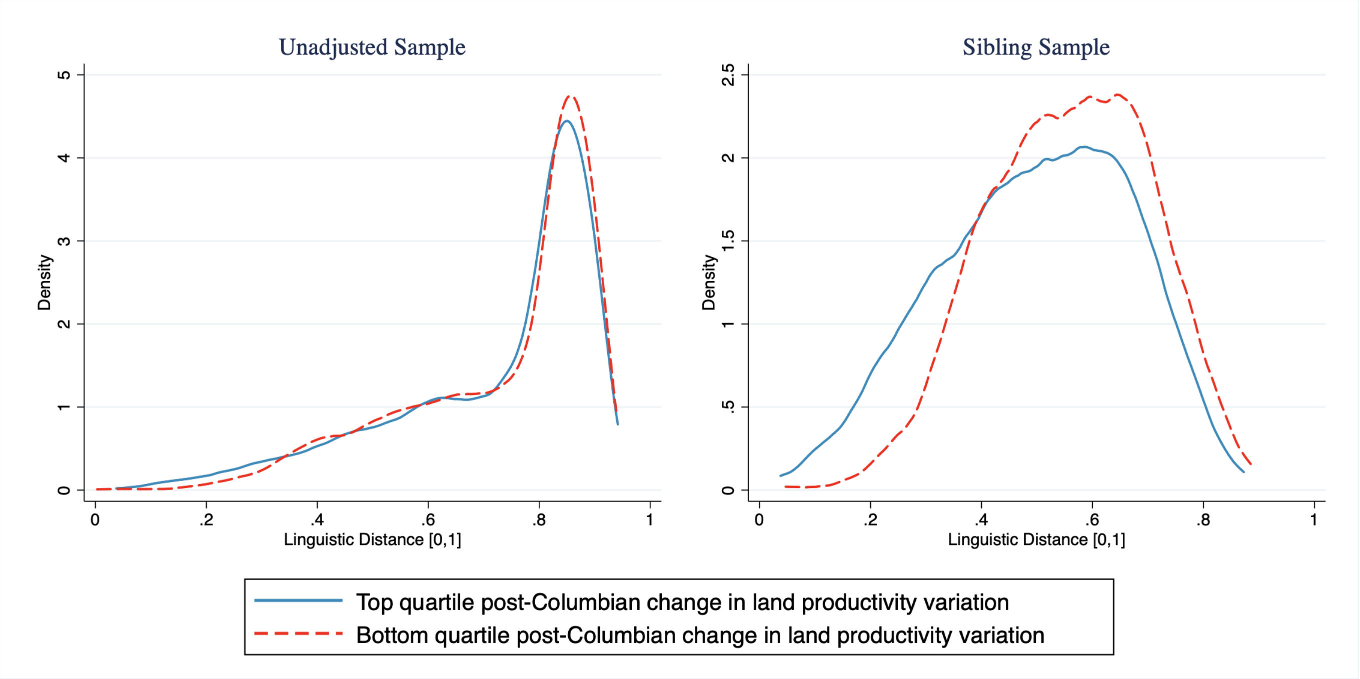
The figure maps the homeland of the Shona and Manyika—spatially adjacent ethnic groups located within the borders of modern-day Zimbabwe. The concurrent segment of border delimiting these group homelands is colour-coded red. A buffer zone is constructed around the segment of shared border, as shown in the figure, that is 100 kilometers in diameter. Throughout the baseline analysis, border buffer zones serve as the unit of observation.

each buffer zone. This measure serves as a proxy for the historical gains from trade between adjacent groups—an assumption that the model in Section 2 makes clear. I do this using Galor and Ozak's (2016) Caloric Suitability Index, a measure of *potential* caloric yields for each $5' \times 5'$ (≈ 100 km^2) grid cell across the globe. My identification strategy relies on the *change* in land productivity variation that results from the Columbian Exchange—the widespread exchange of crops between the Old and New World following Columbus' encounter of the Americas in 1492 (Nunn and Qian, 2010). This unexpected change in the availability of agricultural goods provides a historical source of quasi-random variation in a geographic region's productivity (Galor and Ozak, 2016).

As a measure of cultural similarity, I calculate the linguistic distance between each adjacent group pair using a lexicostatistical measure of distance. Matching languages to groups is straightforward since the *Ethnologue* maps ethnolinguistic groups—ethnic groups unified by a common language. This approach builds on the idea that ethnolinguistic identity is an important predictor of cultural values, norms and preferences (Desmet et al., 2017). The view that language and culture are inseparable is common enough that "language and its associated culture" is considered an adage among ethnolinguists (Risager, 2015, p. 87).

The baseline estimates indicate that a standard deviation increase in land productivity variation at the onset of the Columbian Exchange results in a 2-4 percentage point decrease in

Figure 2: Distribution of Linguistic Distance by Samples



Kernel density plots of linguistic distance for spatially adjacent ethnic group pairs, comparing group pairs in the top and bottom quartiles of post-Columbian changes in land productivity variation. The left figure plots the distribution for the full sample, and the right figure plots the distribution for the ancestry-adjusted sibling sample. All densities are smoothed using a standard Epanechnikov kernel.

linguistic distance.[3] These results are robust to country fixed effects and various geographical confounders, providing clear evidence of the proposed link between land productivity variation and linguistic distance.

Attributing this link to the geographical-trade mechanism of interest is less clear because many factors affect linguistic distance—most notably the ancestral relationship between two groups. I address this concern using phylogenetic data encoded in the *Ethnologue*. Specifically, I narrow my focus to a subset of adjacent *sibling* group pairs that descend from the *same* parent language as a way of disentangling the effect of shared ancestry from the effect of geography that I'm interested in. The sibling-sample findings are qualitatively equivalent to the baseline findings. Yet the estimates are larger in magnitude and are estimated with far more precision, suggesting that identification of the mechanism of interest is indeed clouded by the influence of ancestry on linguistic distance at baseline.

Figure 2 illustrates this main finding by comparing the distribution of linguistic distance for high- and low-variation regions in both the unadjusted and sibling sample. The unadjusted sample shows evidence of the described pattern, though the evidence is weak since shared

[3] To get a sense of the size of these estimates, consider the English word *drink*, which translates to *drinken* in Dutch and *drikke* in Danish. The Dutch translation of *drink* is closer to English, and the baseline effect is roughly equivalent to the increased similarity of English-Dutch relative to English-Danish.

ancestry explains the majority of variation in linguistic distance. Whereas there is clear shift towards lower linguistic distances for high variation regions after adjusting for shared ancestry with the sibling sample.

Large-scale migrations have occurred throughout the post-Columbian era, suggesting that contemporary and historical group locations might differ. I test the sensitivity of my findings to this possible shortcoming with a subsample of group pairs located in countries that were largely unaffected by post-Columbian migrations. In all cases, across all specifications and all samples, the variables of interest remain statistically significant and become larger in magnitude than at baseline. This increase in magnitude is consistent with post-Columbian migrations introducing an attenuation bias due to classical measurement error.

Next I provide evidence in support of the claim that land quality determines a society's productive capabilities in pre-modern times. To do this I use Giuliano and Nunn's (2018) *Ancestral Characteristics of Modern Populations* dataset, which links the *Ethnologue* to pre-colonial group characteristics from Murdock's (1967) *Ethnographic Atlas*. I find that groups located in high-productivity regions rely more on agriculture than groups located in low-productivity regions, who instead rely more on pastoralism and fishing as a means of subsistence. Consistent with the model, these findings suggest that a wider range of tradable goods will be produced in high-variation regions, thus creating an opportunity for specialization and trade.

I also provide direct evidence of the geographical-trade mechanism for a subsample of groups, using additional information from the *Standard Cross-Cultural Sample* (Murdock and White, 1969). I find that, during pre-modern times, groups residing in high-variation regions relied more on trade for food and subsistence than groups residing in low-variation regions. At the heart of this mechanism is the idea that land productivity variation results in more social interactions due to trade. To this end, I show that the custom of marrying outside of one's ethnic group is more common in the high-variation regions where inter-ethnic trade occurred. Yet I also find that inter-ethnic conflict is uncommon in these high-variation regions, suggesting that the link between land productivity variation and linguistic distance is primarily driven by trade and, more generally, peaceful social interactions.

These findings contribute to our understanding of the important role geography plays in the emergence and evolution of ethnic groups. In related work, Michalopoulos (2012) links the spatial distribution of ethnic groups across the globe to variations in geographical factors. He finds that ethnic groups develop human capital suitable to their ecological environment, where such skills are non-transferable across ecological boundaries, resulting in the occurrence of ethnic group boundaries near high-variation geographic regions. Ashraf and Galor (2013b) show that genetic diversity within a population is negatively associated with that population's migratory distance from East Africa, and in related work Ashraf and Galor (2013a) show that these variations in genetic diversity contribute to spatial patterns of ethnic diversity today. Cervellati et al. (2017) and Ahlerup and Olsson (2012) also study the origin and persistence of diversity, highlighting different mechanisms that point to geographic isolation as a driver of ethnic diversity.

However, this evidence only speaks to the extensive margin—why some regions are more diverse than others—and cannot speak to the intensive margin—why we observe different rates of divergence between ancestrally related groups.[4] Yet the impact of these intensive margin differences have contemporary implications for bilateral trade (Melitz, 2008), cross-country income differences (Spolaore and Wacziarg, 2009), idea flows (Dickens, 2018b), the likelihood of international conflict (Spolaore and Wacziarg, 2016), the distribution of public resources (Dickens, 2018a), infant mortality rates (Gomes, 2020) and more (see Spolaore and Wacziarg (2013)). Hence, a subtle yet important question remains unanswered: why are some ethnic groups more dissimilar from each other than others? My principal contribution is evidence that variation in land productivity, through its effect on inter-ethnic trade and social interactions, explains why ancestrally related ethnic groups diverge in language at different rates.

My findings also contribute to our understanding of how geography and history interact and shape patterns of comparative economic development. A prime example of this interaction is found in the Columbian Exchange, where Columbus' arrival in the Americas sparked a widespread exchange of crops, thereby altering the agro-climatic conditions across the Old and New World (Nunn and Qian, 2010). My contribution is evidence that the Columbian Exchange shaped the subsistence activities of pre-industrial ethnic groups and the potential gains from inter-ethnic trade. The long-term implications of this is that the Columbian Exchange altered the co-evolution of language between ethnic groups through its impact on trade and social interactions. These findings complement existing studies that provide important insights into the long-run impact of the Columbian Exchange, such as Nunn and Qian (2011), Galor and Ozak (2016), Iyigun et al. (2017), Galor et al. (2018) and Cherniwchan and Moreno-Cruz (2019).

The rest of the paper is organized as follows. Section 2 outlines a conceptual framework with a simple model that links land productivity variations to inter-ethnic trade. Section 3 describes the data and spatial units of observation used throughout the analysis. Section 4 outlines the empirical model and identification strategy, and presents the baseline results. Section 5 reports evidence in support of the proposed trade mechanism and Section 6 concludes.

2 Conceptual Framework

In a pre-industrial phase of development, agricultural land quality determines a society's productive capabilities. Regions with high variation in land productivity, then, give life to a wide range of producible goods.[5] However, the connection between productivity and production

[4] In a recent manuscript, Blouin and Dyer (2021) make important headway on a related issue. They find that language convergence tends to result from strategic economic interactions, suggesting that language tends to converge towards the society with more economic leverage.

[5] A well-documented characteristic of an ecosystem boundary—the transition zone between various ecosystems—is that boundary regions exhibit high levels of biodiversity (Odum, 1971).

is more complex. Even in uniformly low productivity regions, where few agricultural goods are produced, non-agricultural modes of subsistence such as pastoralism can thrive. Yet pastoralists are often not entirely self-sufficient, since they rely on connections to other groups and regions through trade (Kardulias, 2015). This suggests that, in pre-industrial times, inter-group trade proliferated in regions with a range of producible goods because the economic viability of specialization not only depends on the productive capabilities of the environment, but equally upon the opportunity for exchange with producers of different goods (Bates and Lees, 1977; Bradburd, 1996). Building on this intuition, I develop a Heckscher-Ohlin-style model that links variation in land productivity endowments to inter-group trade in a historical setting.

2.1 The Model

Consider a pre-industrial endowment economy composed of two ethnic groups: one rich (R) and one poor (P). Let there be an agricultural good A and a pastoral good X, where both groups are endowed with the same amount of the pastoral good. The rich group is endowed with $(1+\theta)Y$ of the agricultural good and the poor group is endowed with $(1-\theta)Y$, where Y measures the average productivity of agricultural land and $\theta \in (0,1)$ captures the distribution of productive land. This suggests that a higher θ implies more variation in land quality between groups. The rich group is defined by their larger share of total agricultural output, $(1+\theta)/2$, since total output is equal to $2Y$. Preferences over the two goods are the same for both groups and are given by:

$$U_i = \ln(C_{i,A}) + \beta C_{i,X},$$

where $i = R, P$ and $C_{i,A}$ and $C_{i,X}$ denote consumption of the agricultural and pastoral goods. Let p be the price of the pastoral good in terms of the agricultural good.

Solving each group's utility maximization problem subject to its budget constraint implies the marginal utility from agricultural consumption is equal across groups because they both face the same prices; i.e., $C_{R,A} = C_{P,A} = p/\beta$. This result is due to the log-linear form of group preferences. Similarly, the first-order condition and budget constraint for each group is used to solve for pastoral good consumption as a function of price p:

$$C_{R,X} = \frac{(1+\theta)Y}{p} + X - \frac{1}{\beta},$$
$$C_{P,X} = \frac{(1-\theta)Y}{p} + X - \frac{1}{\beta}.$$

To find the equilibrium price p, equalizing the supply and demand for either good is sufficient. The supply of the agricultural good endowment is $2Y$, and the demand for this good is

$C_{R,A} + C_{P,A} = 2p/\beta$. The market clearing condition yields equilibrium price:

$$p = \beta Y. \tag{1}$$

Substituting (1) into $C_{i,A}$ and $C_{i,X}$ for each group i gives equilibrium consumption values under trade:

$$\begin{aligned} C_{R,A} &= C_{P,A} = Y, \\ C_{R,X} &= X + \frac{\theta}{\beta}, \\ C_{P,X} &= X - \frac{\theta}{\beta}. \end{aligned} \tag{2}$$

I assume that $X > \theta/\beta$ to ensure that $C_{P,X}$ is non-negative.

2.2 Gains from Trade

Here, I consider the level of θ under which both groups prefer trade over autarky. Substituting the equilibrium expressions from (2) into each group's utility function yields utility under trade:

$$\begin{aligned} U_R^{\text{trade}} &= \text{ln}Y + \beta X + \theta, \\ U_P^{\text{trade}} &= \text{ln}Y + \beta X - \theta. \end{aligned}$$

Under autarky, each group i consumes its endowment and earns utility U_i^{autarky}. Hence, the gains from trade for each group are determined as follows:

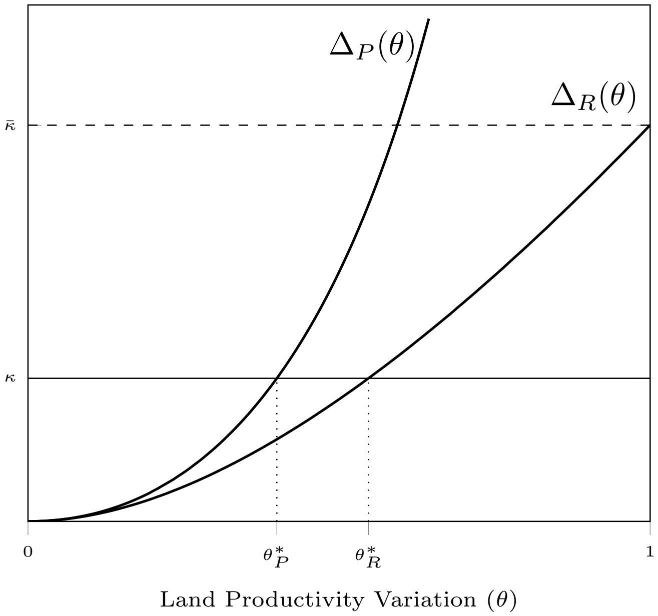
$$\begin{aligned} U_R^{\text{trade}} - U_R^{\text{autarky}} &= \theta - \text{ln}(1+\theta) \equiv \Delta_R(\theta), \\ U_P^{\text{trade}} - U_P^{\text{autarky}} &= -\theta - \text{ln}(1-\theta) \equiv \Delta_P(\theta). \end{aligned} \tag{3}$$

The expressions in (3) imply that the gains from trade are a function of land productivity variation between groups, as depicted in Figure 3. The gains from trade are increasing in θ for both groups, a key prediction of the model and a testable hypothesis that I bring to the data in Section 5. When $\theta \to 1$, log-linear utility implies the poor group's agricultural income goes to zero and they are willing to pay any utility cost for trade. In other words, the poor group is always more willing to trade than the rich group since $\Delta_P(\theta) > \Delta_R(\theta)$ for all $\theta > 0$, as is evident in Figure 3. Yet even at the extreme, where the rich group is endowed with all of the agriculturally productive land ($\theta = 1$), the rich group still prefers trade over autarky as long as the utility cost of trade is less than $\Delta_R(1) = 1 - \text{ln}(2) \equiv \bar{\kappa}$.

Let the utility cost of trade be $\kappa \in (0, \bar{\kappa})$, where the cost is assumed to be positive but modest in size.[6] From Figure 3, it is clear that for some κ there exists a threshold where any $\theta > \theta_R^*$ will

[6] If $\kappa > \bar{\kappa}$, then the cost of trade would be prohibitively high, which is unrealistic since trade was widespread

Figure 3: Gains from Trade



result in trade, since the gains from trade outweigh the utility cost of trade for both groups.

Intuitively, this simple model suggests that the likelihood of trade is greater in regions with large variations in land productivity relative to more homogeneous regions. Historians have noted a similar mechanism of ecologically-driven trade (Lovejoy and Baier, 1975), and economists studying the origin of the state believe that states emerged in regions with ecologically-driven trade to protect growing market economies (Bates, 1983; Fenske, 2014).

2.3 Inter-Ethnic Trade and Linguistic Distance

Historical factors play a significant role in the evolution of culture and language (Cysouw, 2013). The traditional view among linguists is that there are two primary channels of influence: vertical transfer (i.e., genealogical descent) and horizontal transfer (i.e., cross-cultural borrowing). Here, I describe the horizontal nature of the proposed trade mechanism, which underscores the importance of holding constant genealogical differences in the empirical analysis that follows.

It is often theorized that social mechanisms explain bilateral differences in culture (Centola et al., 2007). When two populations engage in collaborative social activities with shared intentions and goals, a cooperative relationship for communication is created (Tomasello, 2008). It is through these cooperative relationships that linguists believe the evolution of culture and language manifests itself and horizontal transfer occurs.

even in the pre-industrial era. Even if transportation costs were high in pre-industrial times, I only consider neighbouring ethnic groups in the empirical analysis, so transport distances were relatively short in the context I consider here.

Historical inter-ethnic trade—a collaborative activity reliant on social interactions—can be understood through this lens, where shared intentions and goals create a cooperative relationship for communication and symbiosis in economic exchange. Trading groups then face an adaptive advantage when their two languages are similar, suggesting that their phylogenetic relationship is influenced by the frequency of inter-ethnic social interactions. Turner et al. (2003, p. 452) reiterate this point, writing that "not only are products of diverse regions and ecosystems shared and redistributed when cultural groups meet and mingle, so too are [. . .] linguistic traits and vocabulary."

Whereas the motives for specialization and trade disappear in homogeneous ecological environments, as equation (3) suggests. The lack of trade and limited social interactions can influence language in many ways, including cultural divergence (Blouin and Dyer, 2021) and an amplification of the natural process of drift (Eggan, 1963; Boyd and Richerson, 1985). Hence, the effect of variation in land productivity, through its effect on inter-ethnic trade and social interactions, explains why some ethnic groups are more similar in language than others.

3 Data

3.1 Border-Level Dataset

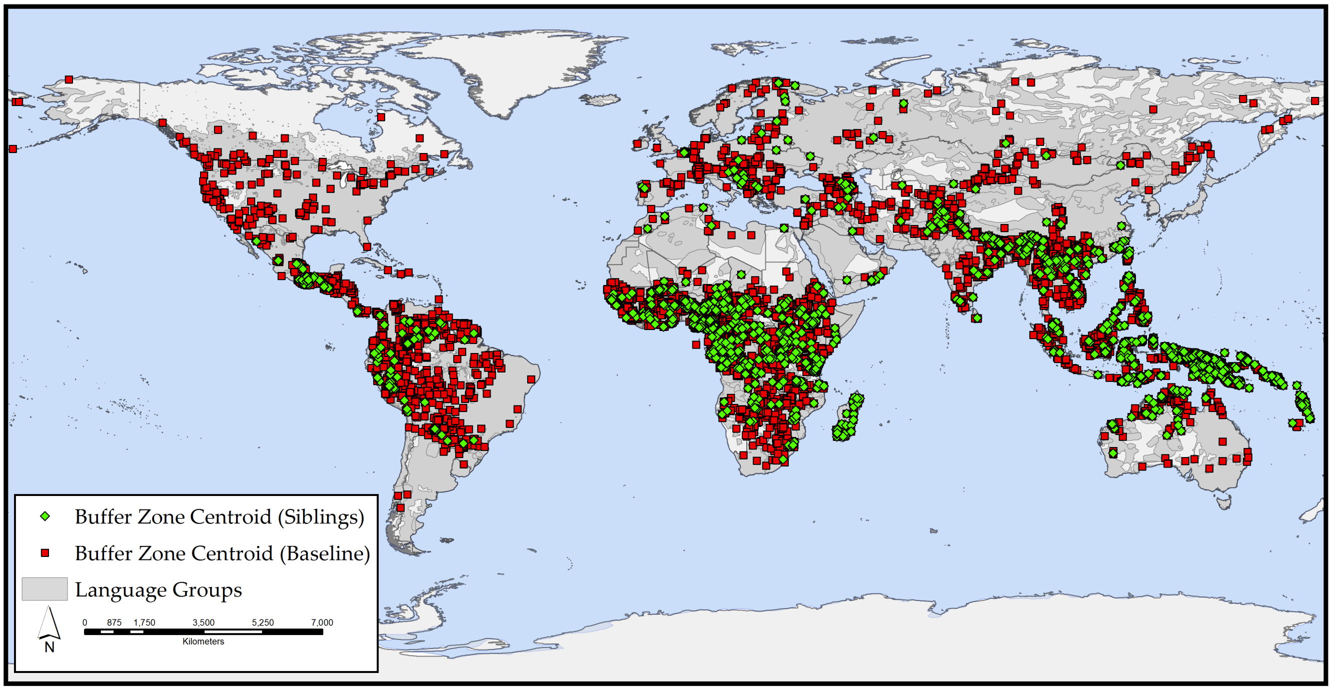
For the baseline analysis, I build a georeferenced dataset where the unit of observation is the segment of border demarcating spatially adjacent ethnic homelands. For this I use data from the *Ethnologue* (Lewis, 2009, 16th edition), including a map of the global distribution of ethnolinguistic groups (WLMS, 2009). The *Ethnologue* maps ethnolinguistic homelands contained within contemporary country borders, implying that the same group may appear in more than one country and receives unique treatment in the data.

To start, I use Geographic Information System (GIS) software to identify all spatially adjacent group pairs across the world with concurrent group-level borders. This amounts to 15,603 unique pairs, where each pair is composed of two groups that each speak a distinct language.[7] I then use GIS to locate the concurrent segment of border delimiting each adjacent pair. Finally, I construct a buffer zone 100 kilometers in diameter around each border segment as my unit of observation.[8] As the unit of observation, these buffer zones serve as a topographic lens to observe the region that links spatially adjacent group pairs.

[7] I limit my search to *Ethnologue* group pairs with reported non-zero populations that are located within the same continent. The initial search identified 17,129 pairs, yet some groups occupy non-adjacent regions of a country, resulting in 1,526 duplicate pairs. I drop these duplicated pairs from the dataset.

[8] More specifically, the GIS software constructs a 50-kilometer radius in every direction for each point along the border segment. The continuous application of this procedure results in a buffer zone that traces the concurrent segment of border with a diameter of 100 kilometers.

Figure 4: Spatial Distribution of Border Buffer Zones


Buffer Zone Centroid (Siblings)
Buffer Zone Centroid (Baseline)
Language Groups
0 875 1,750 3,500 5,250 7,000
Kilometers
N

Dependent Variable: Linguistic Distance

Language is an important aspect of cultural identity that is commonly used by researchers to study the economics of culture (e.g., Ginsburgh and Weber (2016)). Here, I use a computerized lexicostatistical measure of linguistic distance as an estimate of cultural distance. The measure I use was developed by the Automatic Similarity Judgement Program (ASJP), a team of linguists at the Max Planck Institute for Evolutionary Anthropology (Wichmann et al., 2010, 2016).

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. These transcribed lists of words are available for over 60 percent of global languages. For any two languages of interest, I calculate the minimum number of edits necessary to translate the spelling of a word from one language to another using a Levenshtein distance algorithm. A lexicostatistical measure of distance between two languages is calculated as a normalized average of these Levenshtein distances.[9]

For the baseline analysis, I match estimates of linguistic distance to ethnolinguistic group pairs associated with 8,402 buffer zones located in 164 countries. Although this observed sam-

[9] Compared to alternative measures of linguistic distance, the lexicostatistical measure is indispensable to this analysis. Consider the commonly used cladistic measure of linguistic distance between groups—a count of shared nodes on the global language tree. By definition a sibling pair share the maximum number of tree nodes, implying that cladistic distance does not vary among siblings. To the contrary, lexicostatistical distance is a continuous measure of distance that exhibits substantial variation across all sibling pairs. The sibling analysis would not be possible without the additional variation the lexicostatistical measure provides. See Dickens (2018a) for a more in-depth comparison of measures and Online Appendix B for a formal discussion of the estimation procedure.

ple only accounts for 54 percent of the 15,603 identified pairs in the global sample, the overall set of included language groups account for 84 percent of the global population and reside in 89 percent of all countries reported in the *Ethnologue*. In terms of the included language families, the observed sample is quite representative of the global sample too. For example, the three most prominent language families in the global sample (in percentage terms) include Niger-Congo (23 percent), Austronesian (17 percent) and Indo-European (7 percent). In the observed sample, Niger-Congo languages constitute 26 percent, Austronesian 20 percent and Indo-European 8 percent. Figure 4 maps the distribution of the 8,402 buffer zones.

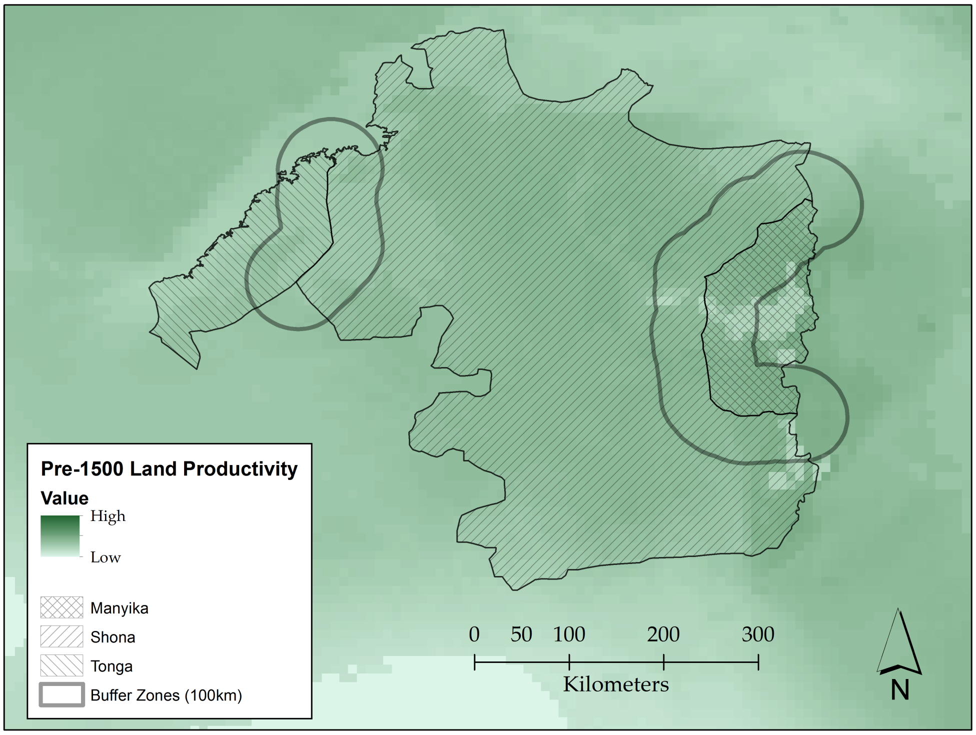
Independent Variable: Land Productivity Variation

A key prediction of the model is that the historical gains from trade are increasing in land productivity variation between groups. Moving from the model to empirics, border buffer zones are a straightforward way to map the geographic terrain shared between groups. To this end, I construct a proxy for the historical gains from trade using Galor and Ozak's (2016) Caloric Suitability Index (CSI) as a measure of buffer zone land productivity.

Data for the CSI measure come from the Global Agro-Ecological Zones (GAEZ) project. Galor and Ozak (2016) use GAEZ crop yield estimates for 48 crops to construct an average measure of potential output for each $5' \times 5'$ grid cell on earth. For each available crop, yield estimates are based on low-level inputs and rain-fed agriculture that reflect traditional labour-intensive cultivation methods used in the distant past. These crop yield estimates are also based on the agro-climatic constraints of a tract of land that are independent of human action, thus reflecting the *potential* output of a grid cell rather than the actual output. These restrictions are important because they address the concern that land quality is potentially endogenous to human intervention. The estimated potential output for each crop is then converted into a standardized measure of potential caloric return. Averaging across all 48 potential yields within each $5' \times 5'$ grid cell results in an average measure of land productivity measured in millions of kilo calories, per hectare, per year.

As a measure of pre-Columbian variation in land productivity, I calculate the standard deviation of pre-1500 CSI grid cells within each buffer zone. I also construct a measure of pre-Columbian land productivity using the mean value of pre-1500 CSI grid cells. Galor and Ozak (2016) make an important distinction between the availability of crops in a grid cell in the pre-1500 CE and post-1500 CE period. The difference between these two periods reflect the change in land productivity that resulted from the expansion of crops in the post-Columbian period. Hence, for my variable of interest, I calculate the change in land productivity variation in the post-Columbian period as the difference between each buffer zone's post-1500 standard deviation and pre-1500 standard deviation. I similarly calculate the change in average land productivity in the post-Columbian period.

Figure 5: Border-Level Analysis: Example Buffer Zones



Two example buffer zones used in the border-level analysis. The Shona ethnic group reside in the large territory in the center of the figure. To the west is their shared border with the Tonga ethnic group and to the east is the Shona's shared border with the Manyika ethnic group. All groups reside within the borders of modern-day Zimbabwe. Underlying this map is pre-1500 CE land productivity data, which can be seen to vary more in the Shona-Manyika buffer zone than the Shona-Tonga buffer zone. The empirical design is based on within-country comparisons analogous to this figure, where the linguistic distance between each group pair is associated with border-level variation in land productivity.

Figure 5 illustrates this approach. I map the homeland of the Shona ethnolinguistic group and two neighboring groups: the Tonga and Manyika. All three groups reside within the contemporary borders of Zimbabwe. Underlying this map is pre-1500 CE land productivity raster data. There is far more observable variation in land productivity within the Shona-Manyika buffer zone to the east (standard deviation = 0.27) than there is in the Shona-Tonga buffer zone to the west (standard deviation = 0.13). The Shona and Manyika are also much more similar in language (26 percent dissimilarity) relative to the Shona and Tonga (69 percent dissimilarity). Although this evidence is only suggestive, the within-country comparison shown here is analogous to the empirical design used throughout Section 4.

Additional Control Variables

I collect a variety of other geographic and climatic data as control variables. Temperature and precipitation data come from WorldClim (2006), which is based on Hijmans et al. (2005). Elevation data comes from the *National Oceanic and Atmospheric Administration* (NOAA, 1999), and ruggedness is calculated as the standard deviation of elevation (Michalopoulos, 2012; Kitamura and Lagerlöf, 2020). As a measure of the disease environment, I use the Malaria Ecology Index from Kiszewski et al. (2004).

I also construct numerous spatial measures. I use data from Natural-Earth (2016) to calculate distances from buffer zone centroids to the nearest coast, and data from NOAA (2017), based on Wessel and Smith (1996), to calculate distances to the nearest lake, major river and minor river. To capture the geographic size of each adjacent group pair, I use GIS to calculate the total land area occupied by both groups and the geodesic distance between group centroids. I also calculate the absolute difference in latitude and longitude of each pair using their centroid coordinates to account for the spatial orientation of an adjacent pair.

Population data comes from the *Ethnologue* (Lewis, 2009). For an adjacent group pair, I add the total population for each group together. I exclude any group pair where one or both of the language groups have a recorded population of zero.[10]

3.2 Ancestral Characteristics Group-Level Dataset

In Section 5, I provide evidence of the geographical-trade mechanism using ethnolinguistic group-level data. This analysis relies mostly on Giuliano and Nunn's (2018) *Ancestral Characteristics of Modern Populations* dataset, which links the *Ethnologue* to pre-colonial group characteristics in Murdock's (1967) *Ethnographic Atlas*. Included with the database is an augmented map of the *Ethnologue*. Here, I use each ethnolinguistic group homeland as the spatial unit of observation, and construct the identical set of land productivity and control variables used in the baseline border-level dataset.

This dataset allows me to measure a variety of historical outcomes at the group level. In particular, I construct a set of indicators denoting each group's historical dependence on agriculture, pastoralism, fishing and hunting-gathering for subsistence. As a measure of inter-group social interactions, I also construct an indicator equal to one if the group is coded as an exogamous community.

I also supplement these data with additional group-level information from Murdock and White's (1969) *Standard Cross-Cultural Sample* (SCCS). These data provide far more details of ancestral characteristics than the *Ethnographic Atlas*, but only for a subset of groups. Importantly, the SCCS encodes a group's reliance on inter-ethnic trade for food and subsistence that I use to test the proposed mechanism. These data also include further information on exogamy, in addition to external conflict—a non-peaceful form of inter-group social interaction.[11]

[10] See Table A1 in the appendix for the full sample and sibling sample summary statistics. Online Appendix C also describes in detail how each variable is created.

[11] See Table A2 in the appendix for summary statistics. Online Appendix C also provides details of how these data were constructed, with reference to the variable number in the *Ethnographic Atlas* and the SCCS.

4 Empirical Strategy and Estimates

4.1 Identification Strategy

The main empirical challenge I face is to connect historical variations in land productivity, through its effect on inter-ethnic trade, to contemporary differences in language. For this, I rely on the natural experiment associated with the Columbian Exchange—the widespread exchange of goods and crops between the Old and New World in the post-1500 period (Nunn and Qian, 2010). Distinct ecological environments across the world were fundamentally changed by the flow of plants and animals in the post-Columbian era (Frankema, 2015). This coming together of the continents introduced a new set of potential crops for cultivation that, in the context of this paper, provides quasi-random variation in potential land productivity.

Michalopoulos (2012) finds that ethnic groups develop human capital suitable to their ecological environment, where such skills are non-transferable across ecological boundaries. This suggests that, in the pre-1500 period, similar groups may have located next to one another due to their relative ease of communication, and defined the frontier of their territories across geographic regions with high variation in land productivity because of their location-specific human capital. While this would still speak to the geographical origins of linguistic distance, the inter-ethnic trade mechanism would be inconsequential.

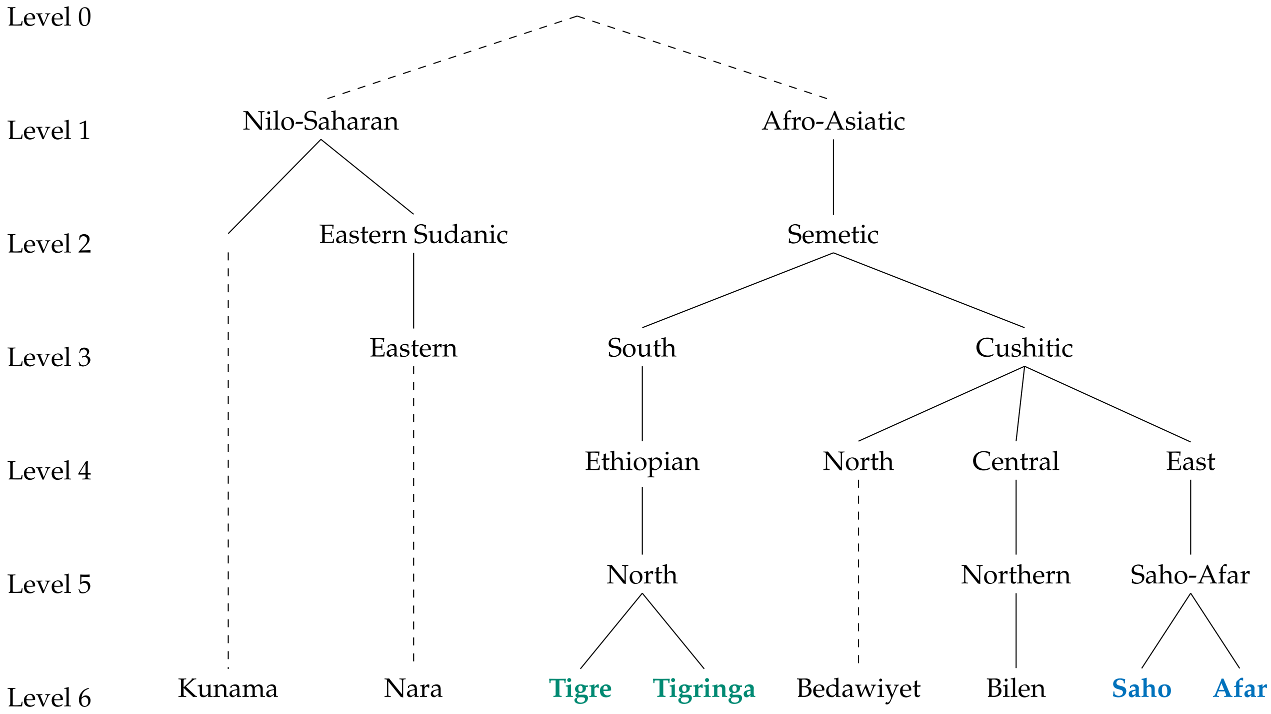
I overcome this concern of endogenous group sorting with quasi-random variation resulting from the Columbian Exchange. The identifying source of variation in land productivity comes from an unexpected change in the potential set of crops for cultivation in the post-Columbian period (Galor and Ozak, 2016). The necessary assumption is that the change in land productivity variation in the post-Columbian period is random and independent of all other determinants of linguistic distance, conditional on the level of land productivity variation in the pre-Columbian period. Hence, this quasi-random variation mitigates concern that similar groups sorted into geographic regions defined by high levels of productivity variation.

The other main empirical challenge I face is disentangling the horizontal transmission of culture via trade from the vertical transfer of culture via genealogical descent. This consideration is important because all related adjacent group pairs exhibit similarity in language due, in part, to genealogical descent.

I overcome this concern by narrowing my focus to sibling ethnolinguistic pairs—those that descend from the same parent language. I define pairs as siblings if they share an identical ancestral history, separated only at the most recent cleavage on the *Ethnologue* phylogenetic tree. For example, Figure 6 depicts the 8 major Eritrean languages for which there are 28 pairings. Only Tigre-Tigringa and Saho-Afar represent sibling pairs, since these pairs share 5 out of 6 branches on the tree—the maximum number of shared branches for two distinct languages.[12]

[12] The *Ethnologue* world tree contains 15 levels, which I abstract from here for simplicity.

Figure 6: Phylogenetic Tree of Eritrean Languages



Example of sibling pairs. This figure depicts the language tree for the 8 major languages of Eritrea. Among the 28 possible pairings of these languages, only 2 represent sibling pairs since there are only 2 language pairs that share a common parent language on the language tree: Tigre-Tigringa and Saho-Afar.

By narrowing my focus to sibling pairs, an equivalent phylogenetic relationship is guaranteed between each pair, thus holding constant the effects of vertical transfer. This implies that sibling sample estimates more reliably identify the horizontal channel of interest, where variation in land productivity determines the extent of trade, and the frequency of contact via trade shapes the transmission of culture.[13]

4.2 Empirical Model and Results

Define buffer zone k as the region surrounding the segment of border that separates ethnolinguistic groups i and j. I estimate the effect of land productivity variation in buffer zone k on the linguistic distance between groups i and j in the following way:

$$LD_k = \beta_0 + \beta^{1500} ProdVar_k + \beta^{change} \Delta ProdVar_k + x_k' \, \Phi + \lambda_{l_i(k)} + \theta_{l_j(k)} + \delta_{c(k)} + \epsilon_k. \qquad (4)$$

The dependent variable LD_k measures the linguistic distance between neighbouring ethnolinguistic groups i and j in buffer zone k. $ProdVar_k$ denotes pre-1500 variation in land pro-

[13] Figure 4 maps the spatial distribution of sibling pairs relative to the baseline sample of group pairs.

ductivity in buffer zone k, and $\Delta ProdVar_k$ denotes the change in land productivity variation in the post-1500 period at the onset of the Columbian Exchange. Here, x_k represents a vector of buffer zone geo-climatic characteristics and spatial control variables.[14] $\lambda_{l_i(k)}$ and $\theta_{l_j(k)}$ respectively denote a complete set of language family fixed effects for groups i and j in buffer zone k, and $\delta_{c(k)}$ is a complete set of country fixed effects associated with buffer zone k. The theoretical framework suggests that $\hat{\beta}^{1500} < 0$ and $\hat{\beta}^{change} < 0$.

4.2.1 Full Sample Baseline Estimates

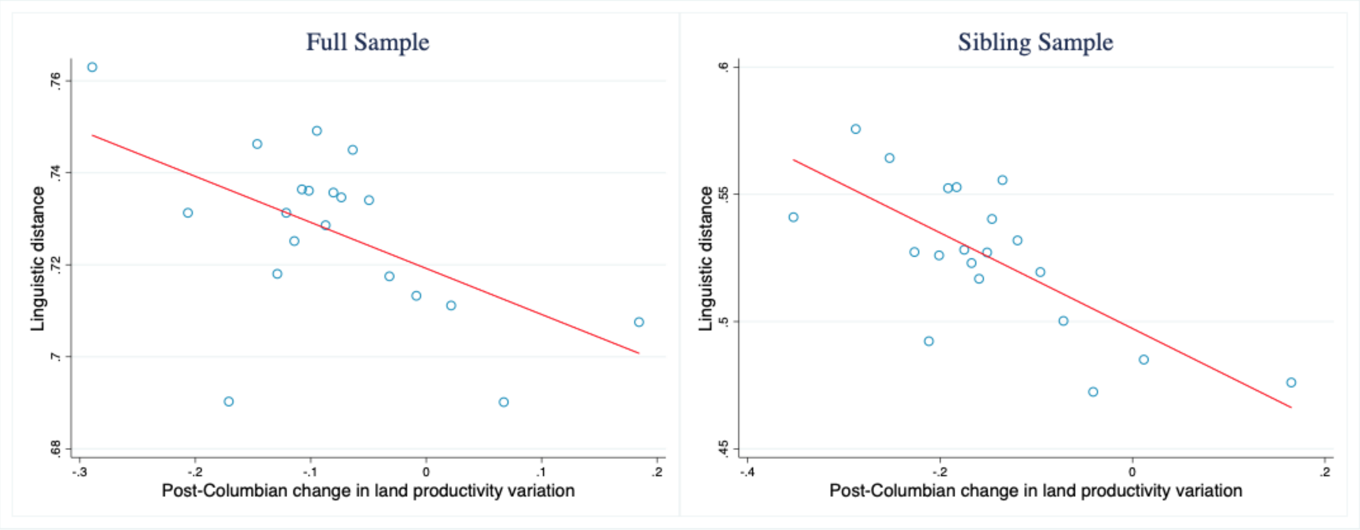
Table 1 presents estimates of equation (4). All reported estimates include language family fixed effects to adjust for deep-rooted ancestral differences in adjacent language pairs.[15] Column 1 reports within-family estimates for pre-Columbian land productivity variation and the post-Columbian change in land productivity variation. Both estimates enter with the expected negative sign and are statistically significant at the 1 percent level. This says that ethnic groups separated across high-variation regions are more similar in language than groups separated across low-variation regions. In particular, a one standard deviation increase in pre-1500 land productivity variation decreases linguistic distance by 1.7 percentage points, while a standard deviation increase in productivity variation at the onset of the Columbian exchanges implies a 2.0 percentage point decrease in linguistic distance. Figure 7 shows this negative relationship is not driven by outliers, using a scatterplot that groups post-Columbian changes in land productivity variation into 20 equal-sized bins.

These baseline findings are robust to adding land productivity controls (column 2), geography controls (column 3) and spatial controls (column 4). The estimates reported in column 5 include the combined set of these control variables. The coefficient on pre-1500 land productivity variation retains the expected sign, but loses significance at standard levels. Whereas the coefficient of interest—the effect of land productivity variation resulting from the Columbian Exchange in the post-1500 period—retains statistical significance with the expected negative

[14] Including pre-1500 land productivity and the change in land productivity in the post-1500 period following the Columbian Exchange; the malaria suitability index; elevation; ruggedness; precipitation and precipitation variation; temperature and temperature variation; log distance between group i and j centroids; log distance to the nearest coast, country border, lake, major river and minor river; the absolute difference in group i and j latitude and longitude coordinates; log total area of an ethnolinguistic pair; and log population.

[15] Figure A2 in the Online Appendix displays estimates of β^{change}, the change in land productivity variation in the post-Columbian period, for 15 different samples with and without language family fixed effects. As described in Section 2, a key empirical challenge I face is to disentangle the horizontal transmission of culture via trade from the vertical transmission of culture via genealogical descent. While the ancestral relationship between each group pair will impact their level of similarity, the vertical nature of genealogical descent within a single group is orthogonal to geographical variations in the surrounding region. Hence, any estimate that excludes language family fixed effects should be biased towards zero by the diminished signal-to-noise ratio. This point is made clear in Figure A2, where $\hat{\beta}^{change}$ is statistically insignificant for the full-sample estimate without fixed effects, but becomes negative and significant after narrowing the sample to increasingly related group pairs. To the contrary, estimates of $\hat{\beta}^{change}$ with language family fixed effects are relatively stable in magnitude, and are always negative and significant irrespective of sample.

Figure 7: Border-Level Scatterplots: Linguistic Distance and Land Productivity Variation



Unit of observation: border buffer zone (100km). Scatterplots grouping post-Columbian changes in land productivity variation into 20 equal-sized bins. The full-sample plot corresponds to column 1 of Table 1 and the sibling-sample plot corresponds to column 1 in Table 2. Both plots are conditional on buffer zone pre-Columbian land productivity variation and language family fixed effects.

sign, and is statistically equivalent to the unconditional estimate in column 1. The estimates reported in column 6 include country fixed effects. The change in land productivity variation retains statistical significance, but is reduced to two-thirds the size of the other baseline estimates. This estimate implies that a standard deviation increase in post-Columbian land productivity variation implies a 1.3 percentage point decrease in linguistic distance.[16]

Figure A1 in the Online Appendix illustrates the balancedness of this baseline sample. I compare high-variation regions to low-variation regions by discretizing post-Columbian changes in land productivity variation at the median. The majority of buffer zone characteristics do not correlate with variation induced by the Columbian Exchange, indicating that the baseline estimates are based off a relatively well-balanced sample. For the handful of variables that do correlate with post-Columbian changes, there is no pattern of characteristic type (e.g., spatial vs. geographical), nor is there a systematic pattern of upward or downward bias.

It is also noteworthy that these estimates are not a byproduct of a buffer zone's size. These baseline findings are robust to reducing the diameter of a buffer zone from 100 kilometers to 50 kilometers, as the results reported in Table A7 of the Online Appendix indicate. In all instances, the variable of interest is negative and statistically significant, and similar in magnitude to the baseline estimates in Table 1.

Overall, the estimated coefficient of interest is stable in magnitude across comparable fixed

[16] See Table A3 in the Online Appendix for a complete table that includes coefficient estimates for all control variables.

Table 1: Border-Level Regressions: Full Sample Baseline Results

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$						
	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	-0.100***	-0.100***	-0.083**	-0.103***	-0.098***	-0.065*
	(0.030)	(0.034)	(0.033)	(0.033)	(0.033)	(0.034)
Land productivity variation (pre-1500)	-0.061***	-0.061***	-0.046*	-0.043**	-0.040	-0.029
	(0.022)	(0.022)	(0.026)	(0.021)	(0.026)	(0.027)
Δ in land productivity (post-1500)		-0.001	0.001	-0.005	-0.001	0.008
		(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Land productivity (pre-1500)		-0.001	-0.001	0.001	0.004	0.009
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Geography Controls	No	No	Yes	No	Yes	Yes
Spatial Controls	No	No	No	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	No	No	Yes
Adjusted R^2	0.25	0.25	0.26	0.26	0.28	0.37
Observations	8402	8402	8402	8402	8402	7291

Unit of observation: border buffer zone (100km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair's lexicostatistical linguistic distance. Geography controls include mean elevation, ruggedness, mean temperature and its standard deviation, mean precipitation and its standard deviation, and the prevalence of malaria. Spatial controls include logged distance to the nearest coast, country border, lake, major river and minor river, logged distance between group centroids, the absolute difference in latitude and longitude, logged land area and logged population. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

effects specifications and various buffer zone sizes. The insensitivity of this estimate to the set of included control variables suggests that the empirical model is well identified, as the balancedness test also suggests. The only source of sensitivity is whether country fixed effects are included or not, but this is to be expected for a variety reasons—e.g., how states integrated different ethnolinguistic groups throughout the historical process of nation building.

4.2.2 Sibling Sample Baseline Estimates

The proposed link between land productivity variation and linguistic distance is horizontal in nature: trade proliferates in high-variation regions, which facilitates social interactions that result in the cross-cultural transmission of language. Yet identification of this channel is not straightforward because the ancestral relationship between groups has a large impact on linguistic differences. By design, the sibling-sample estimates hold genealogical differences constant so the horizontal transmission of language can be identified with greater precision.

Table 2: Border-Level Regressions: Sibling Sample Baseline Results

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0,1)$						
	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	-0.188***	-0.197***	-0.273***	-0.185***	-0.248***	-0.154**
	(0.042)	(0.047)	(0.060)	(0.047)	(0.060)	(0.070)
Land productivity variation (pre-1500)	-0.088**	-0.088**	-0.191***	-0.097***	-0.185***	-0.140**
	(0.034)	(0.034)	(0.059)	(0.035)	(0.059)	(0.068)
Δ in land productivity (post-1500)		-0.007	-0.009	-0.001	-0.003	0.010
		(0.016)	(0.015)	(0.016)	(0.015)	(0.019)
Land productivity (pre-1500)		-0.018	-0.008	-0.018	-0.009	0.011
		(0.012)	(0.011)	(0.012)	(0.012)	(0.014)
Geography Controls	No	No	Yes	No	Yes	Yes
Spatial Controls	No	No	No	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	No	No	Yes
Adjusted R^2	0.28	0.28	0.29	0.30	0.30	0.39
Observations	1669	1669	1669	1669	1669	1497

Unit of observation: border buffer zone (100km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair's lexicostatistical linguistic distance for the baseline sibling sample. Geography and spatial control variables are identical to the complete set of baseline control variables used in Table 1. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

I proceed by reproducing the estimates in Table 1 with the sibling subsample. Table 2 reports these estimates. Both pre-Columbian land productivity variation and the post-Columbian change in land productivity variation are estimated to be negative and statistically significant across all specifications. As expected, the coefficients of interest are estimated with far more precision than at baseline. Figure 7 provides visual evidence of this when comparing binned scatterplots for each sample. The sibling-sample estimates are also larger in magnitude—often twice the size of the comparable baseline estimate. Consider the within-country estimates in column 6, the most demanding specification, where a one standard deviation increase in pre-1500 land productivity variation decreases linguistic distance by 4.4 percentage points, while a standard deviation increase in post-Columbian productivity variation implies a 4.2 percentage point decrease in linguistic distance. At baseline, the comparable numbers were only 0.8 and 1.3 percentage points, respectively. The sibling sample also more balanced than the baseline sample (see Figure A1), adding further credibility to the sibling-sample estimates. Overall, Table 2 provides the strongest evidence of a link between land productivity variations and linguistic distance, and illustrates the quantitative importance of isolating the horizontal transmission channel in this context.[17]

To get a sense of the size of these effects, consider the combined effect of a standard deviation increase in land productivity variation in the pre- and post-Columbian period. The combined effect amounts to an 8.6 percent decrease in linguistic distance, which is roughly equivalent to the relative distance between English and Dutch (64 percent dissimilarity) versus English and Icelandic (73 percent dissimilarity). All three languages are Germanic in origin, however English and Dutch are West Germanic in descent and Icelandic is North Germanic. In other words, the combined baseline estimate of land productivity variation is roughly equivalent to the added linguistic distance of an additional branch on the Indo-European language tree.

4.2.3 Post-Columbian Migrations

The key identifying assumption of the baseline model is that the change in land productivity variation in the post-Columbian period is random and independent of all other factors in a buffer zone, conditional on the level of land productivity variation in the pre-Columbian period. Yet large-scale migrations have occurred throughout the post-Columbian era. This is problematic because the contemporary location of an ethnolinguistic group might differ from their ancestral location, and so historical changes in land productivity would be independent of contemporary language differences.

Historical group-level migration data is unavailable to address this concern. Instead, I minimize the measurement error introduced by post-Columbian migrations by narrowing my focus to adjacent pairs that reside in a country where the majority of people are native to that country. By doing so, I exclude the regions of the world most affected by the wave of post-Columbian migrations (e.g., the Americas).[18] Table A8 in the Online Appendix reports within-country estimates for a variety of specifications, where at least 25, 50 or 75 percent of the population is native to the country of residence, for both the full and sibling samples. Across all specifications, the variables of interest maintain statistical significance and become larger in magnitude than at baseline. This increase in magnitude is consistent with post-Columbian migrations introducing an attenuation bias due to measurement error. Overall, these estimates reassuringly suggest that post-Columbian migrations are inconsequential to the baseline estimates.

4.2.4 Land Homogeneity or Low Land Productivity?

An alternative explanation for my main finding is that linguistic distance is not an outcome of land homogeneity, but rather an outcome of specific groups sorting into regions with uniformly low land productivity in the distant past. Groups who located in regions with uniformly low

[17] See Table A4 in the appendix for a complete table that includes coefficient estimates for all control variables.

[18] The ancestral composition of a country's contemporary population comes from Putterman and Weil (2010).

productivity might have done so due to a lifestyle of subsistence that does not rely on agriculture or inter-group exchange (e.g., hunter-gatherer societies). The distinction made here is important because regions with uniformly low land productivity are by definition low-variation regions, as is evident from the balancedness test in Figure A1. Although I control for pre-Columbian land productivity and the post-Columbian change in land productivity throughout the entire analysis, this only guarantees the trade mechanism holds conditional on the productivity mechanism, rather than ruling it out as a competing explanation.

I proceed here with two tests of the competing mechanism. First, I compare mean differences in linguistic distance for various productivity region types and report these estimates in Table A9 in the Online Appendix. In all instances, across various definitions of region type, I find no significant differences in linguistic distance when comparing group pairs residing in uniformly low productivity regions to groups who are not. The same is true when comparing groups residing in and out of uniformly high productivity regions. Second, I re-estimate the baseline model for the full sample and sibling sample, but drop regions with uniformly low land productivity. Table A10 in the Online Appendix reports these estimates. In all instances, the coefficients of interest increase in magnitude while the standard errors remain relatively constant, resulting in even more precise estimates than at baseline. Overall, these findings rule out this alternative channel and give further credibility to the proposed trade mechanism.

4.2.5 Additional Robustness Checks

In the Online Appendix, I document various other robustness checks. In a Malthusian world, variation in land productivity will correspond to variation in population. To better understand if population size and differences matter in this context, I exclude buffer zones from the analysis that are associated with large aggregate populations and large between-group population differences. Table A11 reports these estimates for the full and sibling sample, and shows that the results are qualitatively similar even after adjusting the sample across these two dimensions.

The *Ethnologue* sometimes maps ethnolinguistic group boundaries as overlapping. This is problematic because a buffer zone will not be uniquely representative of the neighbouring pair. In Table A12, I show that full sample and sibling sample estimates increase in magnitude and significance when excluding overlapping buffer zones from the analysis.

I also explore potential bias due to omitted variables. Table A13 reports Altonji et al. (2005) (AET) statistics on how much stronger selection on unobservables must be compared to selection on observables in order to fully explain the results. Overall, the reported AET statistics suggest the influence of unobservable characteristics must be 1.35 to 4.26 times stronger than the influence of observables to fully account for my baseline findings. I also report Oster's (2019) lower bound for the coefficient of interest, under the most conservative assumption of $R^2_{max} = 1$, and in all instances the coefficient remains negative and sizeable in magnitude.

In Table A14, I report estimates of the baseline model with country fixed effects but allow for spatial correlation in the error term. I find that the coefficient of interest remains statistically significant at spatial correlation cutoffs of 100km, 200km, 300km, 400km and 500km.

I also run a variety of placebo tests, where I limit the sample to geographic regions inhospitable to trade (i.e., sample observations above the 90th percentile in elevation or ruggedness), and New World observations where post-Columbian migrations introduce significant noise to the estimates. The hypothesized trade mechanism is expected to disappear in regions inhospitable to trade, hence breaking the link between land productivity variation and linguistic distance. Whereas any long-run evidence of the trade mechanism should be washed away by significant post-Columbian migration in the New World sample. Table A15 reports the results of these tests, where the baseline result disappears in both the full sample and sibling sample as expected.

5 Mechanism: Inter-Ethnic Trade

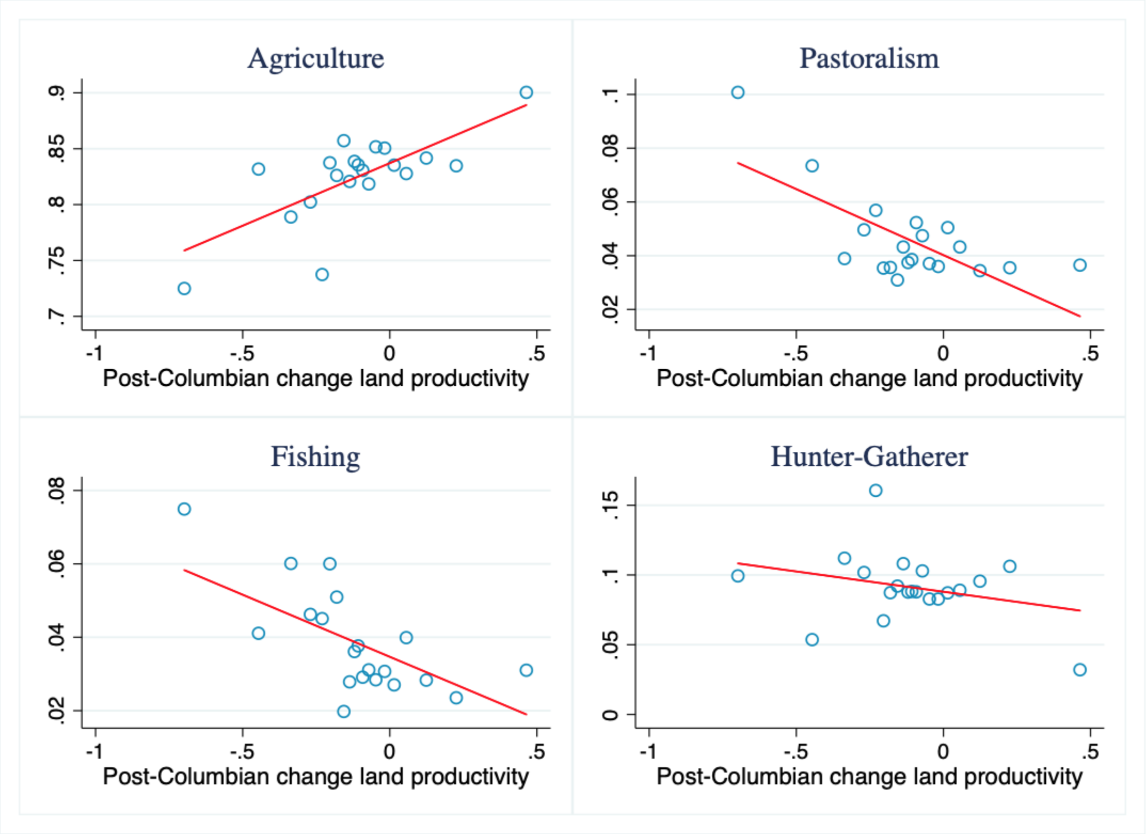
So far, I have established that ethnic groups separated across geographic regions with high variation in land productivity are more similar in language than groups separated across low-variation regions. A framework for interpreting this finding is discussed in Section 2, including the presentation of a simple Heckscher-Ohlin-style model that links variation in land productivity endowments to inter-group trade in a historical setting. In this section, I provide two tests of this mechanism based on the model's predictions: an indirect test using a large sample of pre-colonial ethnic groups, and a direct test using a smaller sample of groups.

5.1 Land Productivity and Historical Modes of Subsistence

Does land productivity predict an ethnic group's historical mode of subsistence? Land productivity should predict subsistence activities, rather than variation in productivity, since land productivity is commonly understood to determine a pre-industrial society's productive capabilities. The extent to which it does is informative because regions with large variations in land productivity are more likely to rely on various modes of subsistence, and thus produce a wider range of tradable goods.

I take this hypothesis to the data using Giuliano and Nunn's (2018) *Ancestral Characteristics* dataset, as described in Section 3. As a first pass of the data, Figure 8 illustrates the relationship between a post-Columbian change in land productivity and four historical subsistence activities, all measured at the ethnolinguistic-group level. These binned scatterplots provide clear evidence of a link between land productivity and a society's primary means of subsistence: agriculture is more common in productive regions whereas non-agricultural subsistence activities are more common in less productive regions.

Figure 8: Group-Level Scatterplots: Historical Subsistence Activity and Land Productivity



Unit of observation: ethnolinguistic group. Scatterplots grouping post-Columbian changes in land productivity into 20 equal-sized bins for various subsistence activities. All plots are conditional on group-level pre-Columbian land productivity, as well as country and language family fixed effects.

Next, I examine this relationship more formally by estimating the following model:

$$Subsistence_i = \alpha_0 + \alpha^{1500} LandProd_i + \alpha^{change} \Delta LandProd_i + x_i' \varphi + \lambda_{l(i)} + \delta_{c(i)} + \varepsilon_i. \qquad (5)$$

The dependent variable $Subsistence_i$ represents one of four subsistence activities: agriculture, pastoralism, fishing or hunting-gathering. In each instance, the indicator takes a value of one if that mode of subsistence is ethnic group i's dominant historical subsistence activity. $LandProd_i$ denotes pre-1500 land productivity within ethnic group i's homeland, and $\Delta LandProd_i$ captures the change in land productivity in the post-Columbian period. x_i denotes a vector of geo-climatic group-level characteristics including pre- and post-Columbian variations in land productivity, and $\lambda_{l(i)}$ and $\delta_{c(i)}$ respectively denote language family and country fixed effects.[19]

Estimates of equation (5) are reported in Table 3. The estimates in column 1 indicate that pre-

[19] Geo-climatic controls include elevation; ruggedness; precipitation and precipitation variation; temperature and temperature variation; log distance to the nearest coast, country border, lake, major river and minor river; the absolute value of latitude and longitude; and log land area. I drop the malaria ecology index to maintain sample size, however the results are qualitatively unchanged by its absence.

Table 3: Group-Level Regressions: Historical Subsistence Activity and Land Productivity

	Dependent Variables: Dominant Historical Subsistence Activity			
	Agriculture	Pastoralism	Fishing	Hunter-Gatherer
	(1)	(2)	(3)	(4)
Δ in land productivity variation (post-1500)	-0.085	0.028	-0.005	0.062
	(0.080)	(0.066)	(0.029)	(0.038)
Land productivity variation (pre-1500)	-0.080	0.010	0.004	0.066
	(0.081)	(0.063)	(0.023)	(0.046)
Δ in land productivity (post-1500)	0.092***	-0.042*	-0.027*	-0.023
	(0.031)	(0.024)	(0.014)	(0.020)
Land productivity (pre-1500)	0.107***	-0.045***	-0.034***	-0.027
	(0.028)	(0.017)	(0.013)	(0.023)
Geography Controls	Yes	Yes	Yes	Yes
Spatial Controls	Yes	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.54	0.37	0.28	0.62
Observations	6128	6128	6128	6128

Unit of observation: ethnolinguistic group. This table illustrates the effect of land productivity on historical modes of subsistence. Each dependent variable is an indicator that takes a value of one if that mode of subsistence is a group's dominant historical subsistence activity. Geography controls include mean elevation, ruggedness, mean temperature and its standard deviation, and mean precipitation and its standard deviation. Spatial controls include logged distance to the nearest coast, country border, lake, major river and minor river, latitude and longitude coordinates, logged land area and logged population. Robust standard errors are clustered at the country level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

colonial groups were more likely to rely on agriculture as a primary means of subsistence when residing in a geographic region with high land productivity. Whereas columns 2 and 3 indicate that pastoralism and fishing were dominant forms of subsistence in low-productivity regions. In other words, low-productivity regions were characterized by non-agricultural subsistence activities. Hunting and gathering societies reside more in low-productivity regions as well, but the estimates in column 4 fall outside of standard levels of confidence.[20]

The results in Table 3 demonstrate that the productive capabilities of a group's homeland predict a group's primary subsistence activity in pre-modern times. Although land productivity variation has no direct impact on the type of subsistence activity, the implication is that high-variation regions are characterized by a variety of subsistence activities. This reasoning aligns with the model, which predicts that the gains from trade are increasing in land productivity variation. Altogether, this evidence supports the inter-ethnic trade mechanism, which is

[20] See Table A5 in the Online Appendix for a complete table that includes coefficient estimates for all control variables.

the first part of the causal chain linking land productivity variations to linguistic differences.

5.2 Land Productivity Variation and Social Interactions

Now I turn to direct evidence of the proposed mechanism with data on inter-ethnic trade and other forms of social interaction. The group-level observation used in the previous subsection is maintained here, but I am limited to a subsample of these groups due to the limited number of groups encoded in the SCCS. To begin, I plot the relationship between the post-Columbian change in land productivity variation and various inter-ethnic social interactions. Consistent with the proposed theory, Figure 9 illustrates that inter-ethic trade is more common in high-variation regions, as is the practice of exogamy. The opposite is true for non-peaceful interactions related to conflict, indicating that the broader link between land productivity variation and linguistic distance is likely driven by social interactions that are peaceful in nature.

I also consider this relationship more formally by estimating the following model:

$$Interaction_i = \mu_0 + \mu^{1500} ProdVar_i + \mu^{change} \Delta ProdVar_i + x_i' \phi + \lambda_{l(i)} + \delta_{c(i)} + \xi_i. \qquad (6)$$

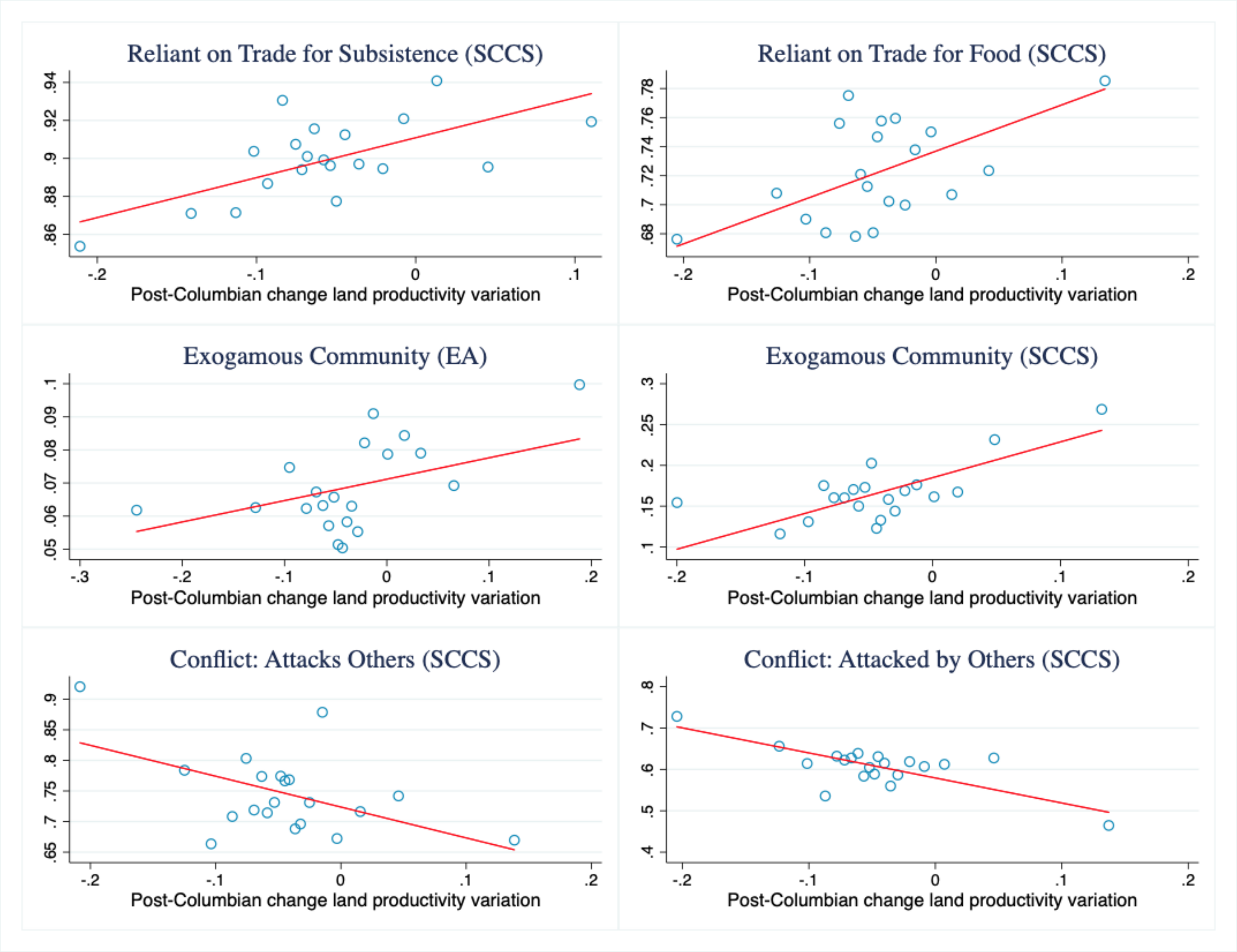
The dependent variable $Interaction_i$ represents one of many indicator variables equal to one if the specified form of inter-ethnic social interaction is present in group i. $ProdVar_k$ denotes pre-1500 land productivity variation in group i, and $\Delta ProdVar_k$ denotes the post-Columbian change in land productivity variation. The geo-climatic control variables and fixed effects are identical to those included in equation (5), as described in the previous subsection.

Table 4 reports estimates of equation (6). Ethnic groups residing in high-variation regions, compared to groups in low-variation regions, are more reliant on inter-ethnic trade as a means of subsistence (column 1) and as a food source (column 2). These findings provide direct empirical evidence of the model's geographical-trade mechanism. The insignificance of average land productivity in both regressions is also consistent with the model's prediction that the historical gains from trade are not a direct outcome of average land productivity.

Next, I provide addition evidence of land productivity variations resulting in inter-ethnic social interactions. Unlike the information on inter-ethnic trade, which is only available in the SCCS, both the *Ethnographic Atlas* (column 3) and the SCCS (column 4) encode measures of exogamy. I find that the custom of marrying outside of one's ethnic group—exogamy—is more common in the high-variation regions where inter-ethnic trade occurred.

Based on the estimates in Table 4, I cannot conclude whether exogamy is an outcome of an existing trade relationship or not. However, the evidence that intermarriage is decreasing in average land productivity likely speaks to a Malthusian mechanism: high-productivity regions can support larger populations than low-productivity regions, and large populations are less reliant on marriage outside of their community, as theory and evidence suggest (Dow et al.,

Figure 9: Group-Level Scatterplots: Social Interactions and Land Productivity Variation



Unit of observation: ethnolinguistic group. Scatterplots grouping post-Columbian changes in land productivity variation into 20 equal-sized bins for various inter-ethnic social interactions. All plots are conditional on group-level pre-Columbian land productivity variation, as well as country and language family fixed effects.

2016).[21] Yet the relationship between the post-Columbian change in land productivity variation and exogamy holds conditional on average productivity—i.e., the relationship of interest holds in addition to the Malthusian mechanism. This is at least suggestive that inter-ethnic trade relationships influence the practice of exogamy through a positive feedback loop of repeated social interactions, thus further influencing the horizontal transmission of language.

Finally, I consider the role of conflict—a non-peaceful form of social interaction that shapes the co-evolution of language.[22] Conflict over unequal land endowments is plausible, and the resulting population exchanges due to conflict would presumably reduce the linguistic differences between warring groups. This consideration is important given a long history of conflict

[21] A similar Malthusian mechanism is plausible with respect to trade, but the insignificance of land productivity in columns 1 and 2 suggest otherwise.

[22] For example, the Magyar invaded Hungary in the ninth century, imposing their Uralic language on the conquered (Cavalli-Sforza, 2000).

Table 4: Group-Level Regressions: Social Interactions and Land Productivity Variation

	Dependent Variables: Social Interactions					
	Inter-Ethnic Trade		Inter-Ethnic Marriage		Inter-Ethnic Conflict	
	For Subsistence	For Food	Exogamy (EA Sample)	Exogamy (SCCS Sample)	Frequently Attacks Others	Frequently Is Attacked
	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	0.201**	0.316**	0.114**	0.375*	-0.502**	-0.578**
	(0.097)	(0.157)	(0.056)	(0.199)	(0.218)	(0.232)
Land productivity variation (pre-1500)	0.139*	0.165	0.072	0.260	-0.349*	-0.472**
	(0.079)	(0.105)	(0.052)	(0.165)	(0.183)	(0.216)
Δ in land productivity (post-1500)	0.022	0.128	-0.005	-0.167**	0.005	0.045
	(0.025)	(0.084)	(0.020)	(0.083)	(0.068)	(0.069)
Land productivity (pre-1500)	-0.002	0.063	-0.038*	-0.113**	-0.037	0.006
	(0.019)	(0.046)	(0.021)	(0.052)	(0.045)	(0.052)
Geography Controls	Yes	Yes	Yes	Yes	Yes	Yes
Spatial Controls	Yes	Yes	Yes	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.90	0.75	0.29	0.68	0.66	0.78
Observations	688	1096	5458	1161	899	869

Unit of observation: ethnolinguistic group. This table illustrates the effect of land productivity variations on inter-ethnic trade and other forms of social interactions. Each dependent variable is an indicator that takes a value of one for group's who engage in the defined inter-ethnic social interaction. Geography controls include mean elevation, ruggedness, mean temperature and its standard deviation, and mean precipitation and its standard deviation. Spatial controls include logged distance to the nearest coast, country border, lake, major river and minor river, latitude and longitude coordinates, logged land area and logged population. Robust standard errors are clustered at the country level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

between agriculturalists and pastoralists (McGuirk and Nunn, 2020).

In Table 4, I report estimates that distinguish between groups who frequently attack others (column 5) and groups who are frequently attacked by others (column 6). In both instances, land productivity variation actually reduces the likelihood of external conflict. This evidence rules out the suggestion that non-peaceful interaction through conflict is an alternative explanation for my baseline findings. A likely explanation for this result is that a peaceful coexistence can be sustained when complementary activities such as inter-ethnic trade and intermarriage exist and are difficult to replicate otherwise (Jha, 2013, 2018). In other words, pre-existing trade relationships can foster a legacy of inter-ethnic tolerance.[23]

5.3 Discussion

The empirical evidence described throughout this section sheds light on the link between land productivity variation and linguistic distance. The estimates reported in Table 3 show that historical group-level productive capabilities are an outcome of geographic land endowments,

[23] See Table A6 in the Online Appendix for a complete table that includes coefficient estimates for all control variables.

and the resulting implication that variation in these endowments give life to specialization and trade is supported by theory (Section 2) and evidence (Table 4).

The evidence in Table 4 also suggests that inter-ethnic trade can foster relationships that extend to norms of intermarriage. Although the direction of causality between trade and intermarriage is speculative, the existence of a link nonetheless verifies that regions with high variation in land productivity are characterized by a variety of inter-ethnic social interactions of a peaceful nature—i.e., the cooperative social dynamics that linguistics believe to be the basis for the horizontal transmission of language and culture across groups (Tomasello, 2008).

Overall, the evidence is strongly supportive of trade as a key mechanism linking land productivity variations to linguistic differences. Table A16 in the appendix provides additional support for this link. I find that adjacent ethnic groups located near Old World trade routes are more similar in language today than neighbouring groups located further away. While these estimates do not rely on post-Columbian changes, and thus cannot be interpreted as causal, the evidence is suggestive of the proposed mechanism and consistent with the idea that groups living near historical trade routes engaged in trade more often than groups living far away (Michalopoulos et al., 2016, 2018).

6 Concluding Remarks

In this study, I take the economic importance of ethnic group differences as given, and go a step deeper to explore the important role geography and trade play in the emergence and co-evolution of these groups. I construct a novel georeferenced dataset to examine the border region of neighbouring ethnolinguistic groups, together with variation in the set of potentially cultivatable crops at the onset of the Columbian Exchange, to estimate how variation in land productivity impacts linguistic differences between neighbouring groups. I find that neighbouring ethnic groups that are separated across regions with more heterogeneous land productivities speak more similar languages.

I argue that inter-ethnic trade is the causal chain connecting geographic variations to linguistic variations, and provide a simple model to clarify the proposed mechanism. Using a variety of data and methods, I find strong support for this mechanism, including direct evidence of a causal link between land productivity variation and a group's reliance on trade for food and subsistence in pre-modern times.

What does this result add to our understanding of the link between ethnolinguistic differences and contemporary patterns of development? It implies that other findings that have been interpreted as effects of ethnolinguistic distance might be rooted in geography. Given the evidence of geography's influence over comparative economic development, my findings suggest that the exogeneity of ethnolinguistic distance in regression analysis should be questioned in the absence of the appropriate geographic control variables.

These findings also give new perspective to how culture evolves. Cultural groups are adaptive to the geographic environment they inhabit and develop location-specific human capital skills (Michalopoulos, 2012). The persistence of cultural traits is similarly dependent on the surrounding environment—during periods of climatic stability, location-specific traits persist within groups because the "rules of thumb" of past generations remain relevant to the current generation (Giuliano and Nunn, 2020). By the same token, the vertical transmission of location-specific skills from parent to child will occur less frequently when the geographic environment is highly variable. Yet I find that, in variable environments, inter-group social interactions are more frequent and these interactions provide a basis for the horizontal transmission of culture. Taken together, this evidence suggests that a cultural group's surrounding geographic environment will influence the extent to which group members pass along existing traits to the next generation or adopt new traits from outside groups. The relative contribution of these mechanisms is an interesting avenue of future research, particularly in this era of climate change where past norms may become less relevant with changing environmental conditions.

References

Ahlerup, P. and Olsson, O. (2012). The Roots of Ethnic Diversity. *Journal of Economic Growth*, 17(2):71–102.

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the Origins of Gender Roles: Women and the Plough. *The Quarterly Journal of Economics*, 128(2):469–530.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy*, 113(1):151–184.

Ashraf, Q. and Galor, O. (2013a). Genetic Diversity and the Origins of Cultural Fragmentation. *American Economic Review: Papers & Proceedings*, 103(3):528–533.

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

Bates, D. G. and Lees, S. H. (1977). The Role of Exchange in Productive Specialization. *American Anthropologist*, 27:824–841.

Bates, R. H. (1983). *Essays on the Political Economy of Rural Africa*. Cambridge University Press, Cambridge.

Bates, R. H. (2010). *Prosperity and Violence: The Political Economy of Development*. W. W. Norton & Company, New York.

Becker, S. O., Boeckh, K., Hainz, C., and Woessmann, L. (2016). The Empire Is Dead, Long Live the Empire! Long-Run Persistence of Trust and Corruption in the Bureaucracy. *Economic Journal*, 126(590):40–74.

Blouin, A. and Dyer, J. (2021). How Cultures Converge: An Empirical Investigation of Trade and Linguistic Exchange. *University of Toronto, Unpublished Manuscript*.

Boyd, R. and Richerson, P. J. (1985). *Culture and the Evolutionary Process*. University of Chicago Press, Chicago.

Bradburd, D. A. (1996). Toward an Understanding of the Economics of Pastoralism: The Balance of Exchange Between Pastoralists and Nonpastoralists in Western Iran, 1815-1975. *Human Ecology*, 24(1):1–38.

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

Centola, D., Gonzalez-Avella, J. C., Eguiluz, V. M., and Miguel, M. S. (2007). Homophily, Cultural Drift, and the Co-Evolution of Cultural Groups. *Journal of Conflict Resolution*, 51(6):905–929.

Cervellati, M., Chiovelli, G., and Esposito, E. (2017). Bite and Divide: Ancestral Exposure to Malaria and the Emergence and Persistence of Ethnic Diversity in Africa. *University of Bologna, Unpublished Manuscript*.

Chanda, A., Justin Cook, C., and Putterman, L. (2014). Persistence of Fortune: Accounting for Population Movements, There Was No Post-Columbian Reversal. *American Economic Journal: Macroeconomics*, 6(3):1–28.

Cherniwchan, J. and Moreno-Cruz, J. (2019). Maize and Precolonial Africa. *Journal of Development Economics*, 136:137–150.

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.

Cysouw, M. (2013). Disentangling Geography from Genealogy. In *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*, pages 21–37. de Gruyter, Berlin.

Desmet, K., Ortuño-ortín, I., and Wacziarg, R. (2017). Culture, Ethnicity, and Diversity. *American Economic Review*, 107(9):2479–2513.

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

Dickens, A. (2018b). Population Relatedness and Cross-Country Idea Flows: Evidence from Book Translations. *Journal of Economic Growth*, 23(4):367–386.

Dow, G. K., Reed, C. G., and Woodcock, S. (2016). The Economics of Exogamous Marriage in Small-Scale Societies. *Economic Inquiry*, 54(4):1805–1823.

Eggan, F. (1963). Cultural Drift and Social Change. *Current Anthropology*, 4(4):347–355.

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

Fenske, J. (2014). Ecology, Trade, and States in Pre-Colonial Africa. *Journal of the European Economic Association*, 12(3):612–640.

Frankema, E. (2015). The Biogeographic Roots of World Inequality: Animals, Disease, and Human Settlement Patterns in Africa and the Americas Before 1492. *World Development*, 70(016):274–285.

Galor, O. and Ozak, O. (2016). The Agricultural Origins of Time Preference. *American Economic Review*, 106(10):3064–3103.

Galor, O., Ozak, O., and Sarid, A. (2018). Geographical Roots of the Coevolution of Cultural and Linguistic Traits. *NBER Working Paper 25289*.

Ginsburgh, V. A. and Weber, S. (2016). Linguistic Distances and Ethnolinguistic Fractionalization and Disenfranchisement Indices. In Ginsburgh, V. A. and Weber, S., editors, *The Palgrave Handbook of Economics and Language*, pages 137–173. Palgrave Macmillan UK, London.

Giuliano, P. and Nunn, N. (2013). The Transmission of Democracy: From the Village to the Nation-State. *American Economic Review: Papers & Proceedings*, 103(3):86–92.

Giuliano, P. and Nunn, N. (2018). Ancestral Characteristics of Modern Populations. *Economic History of Developing Regions*, 33(1):1–17.

Giuliano, P. and Nunn, N. (2020). Understanding Cultural Persistence and Change. *The Review of Economic Studies*, page Forthcoming.

Gomes, J. F. (2020). The Health Costs of Ethnic Distance: Evidence from Sub-Saharan Africa. *Journal of Economic Growth*, 25(2):195–226.

Guiso, L., Sapienza, P., and Zingales, L. (2016). Long-Term Persistence. *Journal of the European Economic Association*, 14(6):1401–1436.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very High Resolution Interpolated Climate Surfaces for Global Land Areas. *International Journal of Climatology*, 25(15):1965–1978.

Iyigun, M., Nunn, N., and Qian, N. (2017). The Long-run Effects of Agricultural Productivity on Conflict, 1400-1900. *NBER Working Paper 24066*.

Jha, S. (2013). Trade, Institutions, and Ethnic Tolerance: Evidence from South Asia. *American Political Science Review*, 107(4):806–832.

Jha, S. (2018). Trading for Peace. *Economic Policy*, 33(95):485–526.

Kardulias, P. N. (2015). Introduction: Pastoralism as an Adaptive Strategy. In Kardulias, P. N., editor, *The Ecology of Pastoralism*. University Press of Colorado, Boulder.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., and Sachs, J. (2004). A Global Index Representing the Stability of Malaria Transmission. *American Journal of Tropical Medicine and Hygiene*, 70(5):486–498.

Kitamura, S. and Lagerlöf, N.-P. (2020). Geography and State Fragmentation. *Journal of the European Economic Association*, 18(4):1726–1769.

Koetsier, T. (2019). From Early Trade and Communication Networks to the Internet, the Internet of Things and the Global Intelligent Machine (GIM). *Advances in Historical Studies*, 08(01):1–23.

Lewis, M. P. (2009). *Ethnologue: Languages of the World (Sixteenth Edition)*. SIL International, Dallas, Texas.

Lovejoy, P. E. and Baier, S. (1975). The Desert-Side Economy of Central Sudan. *The International Journal of African Historical Studies*, 8(4):551–581.

McGuirk, E. F. and Nunn, N. (2020). Nomadic Pastoralism, Climate Change and Conflict in Africa. *NBER Working Paper 28243*.

Melitz, J. (2008). Language and Foreign Trade. *European Economic Review*, 52(4):667–699.

Michalopoulos, S. (2012). The Origins of Ethnolinguistic Diversity. *American Economic Review*, 102(4):1508–1539.

Michalopoulos, S., Naghavi, A., and Prarolo, G. (2016). Islam, Inequality and Pre-Industrial Comparative Development. *Journal of Development Economics*, 120:86–98.

Michalopoulos, S., Naghavi, A., and Prarolo, G. (2018). Trade and Geography in the Spread of Islam. *Economic Journal*, 128(616):3210–3241.

Murdock, G. P. (1967). *Ethnographic Atlas*. University of Pittsburgh Press, Pittsburgh.

Murdock, G. P. and White, D. R. (1969). Standard Cross-Cultural Sample. *Ethnology*, 9:329–369.

Natural-Earth (2016). Natural Earth 1:10m Physical Vectors Coastline.

NOAA (1999). *The Global Land One-Kilometer Base Elevation (GLOBE) Digital Elevation Model (Version 1)*. National Oceanic and Atmospheric Administration, National Geophysical Data Center, Boulder.

NOAA (2017). *Global Self-Consistent, Hierarchical, High-Resolution Geography Database (GSHHG) Version 2.3.7*. National Oceanic and Atmospheric Administration, National Geophysical Data Center, Boulder.

Nunn, N. (2012). Culture and the Historical Process. *Economic History of Developing Regions*, 27(S1):S108–S126.

Nunn, N. and Qian, N. (2010). The Columbian Exchange: A History of Disease, Food, and Ideas. *Journal of Economic Perspectives*, 24(2):163–188.

Nunn, N. and Qian, N. (2011). The Potato's Contribution to Population and Urbanization: Evidence From A Historical Experiment. *The Quarterly Journal of Economics*, 126(2):593–650.

Odum, E. P. (1971). *Fundamentals of Ecology*. Philadelphia, 3rd edition.

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business Economics and Statistics*, 37(2):187–204.

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.

Risager, K. (2015). Linguaculture: the Language-Culture Nexus in Transnational Perspective. In Sharifian, F., editor, *The Routeledge Handbook of Language and Culture*, pages 87–99. New York.

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. (2016). War and Relatedness. *Review of Economics and Statistics*, 98(5):925–939.

Tomasello, M. (2008). *Origins of Human Communication*. The MIT Press, Cambridge.

Turner, N. J., Davidson-Hunt, I. J., and O'Flaherty, M. (2003). Living on the Edge: Ecological and Cultural Edges as Sources of Diversity for Social-Ecological Resilience. *Human Ecology*, 31(3):439–461.

Voigtlander, N. and Voth, H.-J. (2012). Persecuation Perpetuated: The Medieveal Origins of Anti-Semitic Violence in Nazi Germany. *Quarterly Journal of Economics*, 127(3):1339–1392.

Wessel, P. and Smith, W. H. F. (1996). A Global Self-consistent, Hierarchical, High-resolution Shoreline Database. *Journal of Geophysical Research*, 101(B4):8741–8743.

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

Wichmann, S., Holman, E. W., and Brown, C. H. (2016). *The ASJP Database (Version 17)*.

WLMS (2009). *World Language Mapping System (Version 16)*. SIL International, Dallas, Texas.

WorldClim (2006). *WorldClim 1.4 Climate Data for 1960-1990*.

Understanding Ethnolinguistic Differences: The Roles of Geography and Trade

ONLINE APPENDIX

Andrew Dickens[†]

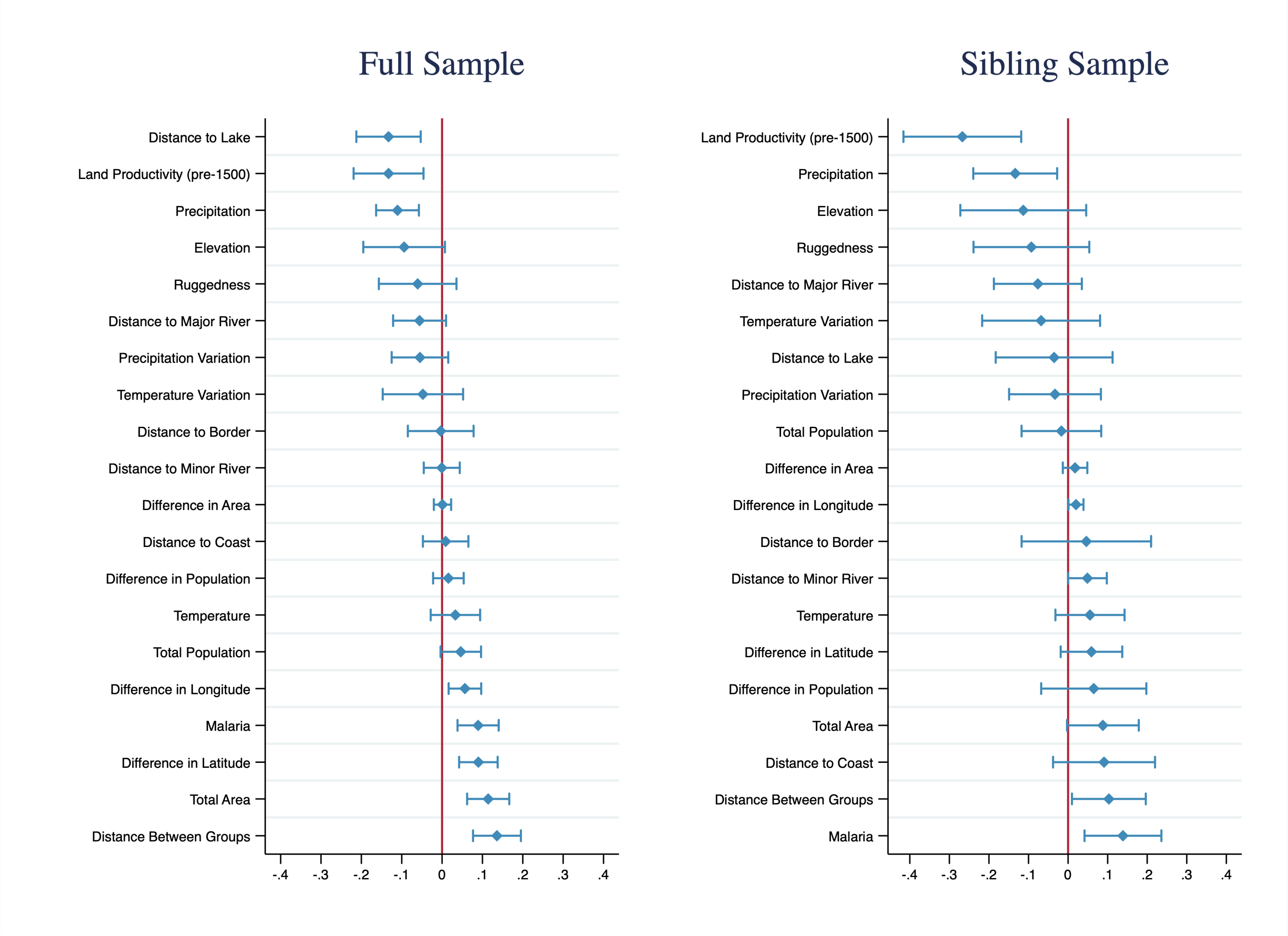
For publication in the

The Economic Journal

[†] Brock University, Department of Economics, St. Catharines, ON, Canada. E-mail: adickens@brocku.ca.

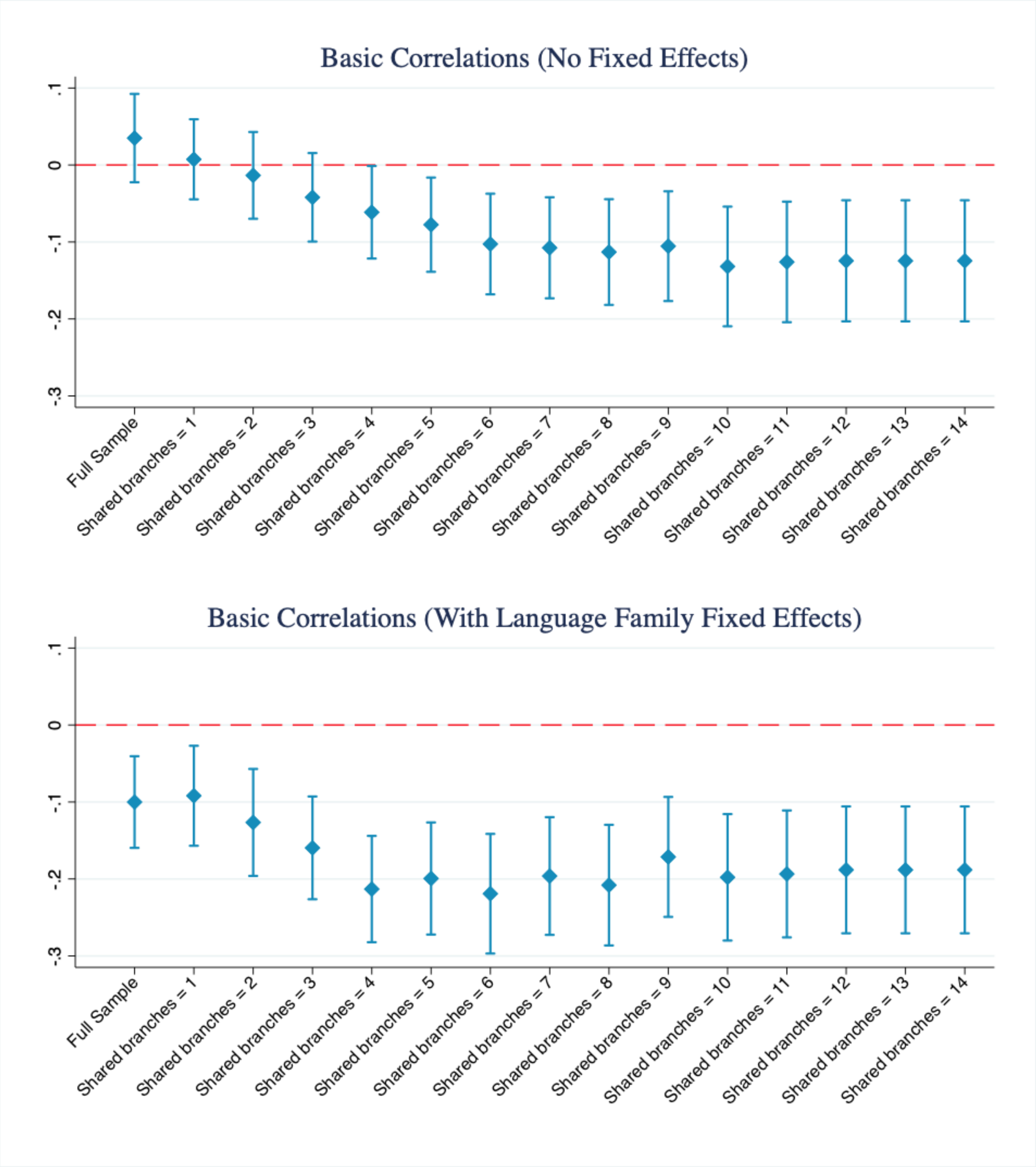
A Supplementary Material

Figure A1: Balancedness



Unit of observation: border buffer zone (100km). This figure illustrates balancedness of the full sample and sibling sample. To implement this test, I use the median of the post-Columbian change in land productivity variation to distinguish high-variation regions from low-variation regions. Every point estimate and confidence interval represent a separate regression of the mentioned variable on the binary treatment variable. Language family and country fixed effects are included in each regression, and standard errors are double-clustered at the level of each language group. Intervals reflect 95% confidence levels.

Figure A2: Linguistic Distance and Post-1500 Change in Land Productivity Variation



Unit of observation: border buffer zone (100km). This figure depicts unconditional estimates of β^{change} from equation (4) in the paper—the change in land productivity variation in the post-Columbian period—for 15 different samples with and without language family fixed effects. The dependent variable for each regression is linguistic distance, and the only included covariate is pre-Columbian land productivity variation. The full sample estimates come from the baseline sample, and, when moving from left to right, the subsequent estimates come from increasingly smaller subsamples of adjacent language pairs who share the stated number of branches on the global *Ethnologue* language tree. The most restricted sample, with 14 shared branches, represents the sibling sample used throughout much of the empirical analysis. Intervals reflect 95% confidence levels.

Units of Observation and Summary Statistics

Border-Level Analysis—The spatial unit of observation a border buffer zone. Each buffer zone is based on the segment of border delimiting a spatially adjacent ethnolinguistic group in the *Ethnologue* (Lewis, 2009, 16th edition). I use GIS software to construct a buffer zone that is 100 kilometers in diameter. To do this, the software constructs 50-kilometer radius in every direction for each point along the border segment—i.e., the radius extends 50 kilometers into the homeland of each group. The continuous application of this procedure results in a buffer zone that traces the concurrent segment of border with an overall diameter of 100 kilometers.

Table A1 reports summary statistics for all variables measured at the level of the buffer zone. Panel A reports summary statistics for the 8,402 border buffer zones used in the full-sample analysis, and Panel B reports summary statistics for the 1,669 buffer zones used in the sibling-sample analysis. The results of both analyses are discussed in Section 4 of the paper.

Group-Level Analysis—The spatial unit of observation is an ethnolinguistic group in Giuliano and Nunn (2018) *Ancestral Characteristics of Modern Populations* dataset. In particular, I use the augmented *Ethnologue* map that comes included with these data. I use the expanded version of this dataset, which supplements details from the *Ethnographic Atlas* with the peoples of Eastern European (Bondarenko et al., 2005) and Siberian (Korotayev et al., 2004), and additional ethnic groups from the *World Ethnographic Sample*. See Giuliano and Nunn (2018) for more information on these extended samples. I also supplement these data with additional information from Murdock and White (1969) *Standard Cross-Cultural Sample*.

Table A2 reports summary statistics for all variables measured at the level of the ethnolinguistic group. The results of the analysis using these data are discussed in Section 5 of the paper.

Main Empirical Evidence Including Control Variable Estimates

Border-Level Analysis—Table A3 is a replication of Table 1, but includes the coefficient estimates for all control variables. Similarly, Table A4 is a replication of Table 2 that includes coefficient estimates for all control variables. The abbreviated version of these tables are displayed in the main text.

Group-Level Analysis—Table A5 is a replication of Table 3, but includes the coefficient estimates for all control variables. Similarly, Table A6 is a replication of Table 4 that includes coefficient estimates for all control variables. The abbreviated version of these tables were displayed in the main text.

Border-Level Regressions: Additional Robustness Checks

Alternative Buffer Zone Size—The baseline estimates are obtained using an arbitrarily-drawn buffer zone of 100 kilometers in diameter. Here, I test the robustness of the baseline estimates using a 50-kilometer border buffer zone. Table A7 reports these estimates. In all instances, the variable of interest is negative and statistically significant, and similar in magnitude to the baseline estimates in Table 1. Overall, this suggests the baseline result is not an outcome of the arbitrarily-sized buffer zones.

Post-Columbian Migrations—The key identifying assumption of the baseline model is that the change in land productivity variation in the post-Columbian period is random and independent of all other factors in a buffer zone, conditional on the level of land productivity variation in the pre-Columbian period. Yet large-scale migrations have occurred throughout the post-Columbian era. This is problematic because the contemporary location of a group might differ from their ancestral location, and so historical changes in land productivity would be independent of contemporary language differences.

I test to robustness of the baseline estimates using various subsamples of groups who reside in countries largely unaffected by post-Columbian migrations. First, I determine the fraction of a contemporary country's population that is native to that country using the World Migration Matrix (Putterman and Weil, 2010). Larger native populations imply less exposure to post-Columbian migrations. Second, I construct three indicator variables equal to one if a buffer zone is located within a country where *at least* (i) 25 percent (ii) 50 percent or (iii) 75 percent of the population is native to the country of residence. Third and finally, I re-estimate my baseline model for subsamples where at least 25, 50 or 75 percent of the population is native to the country of residence, for both the full and sibling sample. Table A8 reports these results. Across all specifications, the variables of interest maintain statistical significance and become larger in magnitude than at baseline. This increase in magnitude is consistent with post-Columbian migrations introducing an attenuation bias due to measurement error. All together, these estimates reassuringly suggest that post-Columbian migrations are inconsequential to the baseline estimates.

Land Homogeneity or Low Land Productivity?—An alternative explanation for my main finding is that linguistic distance is not an outcome of land homogeneity, but rather an outcome of specific groups sorting into regions with uniformly low land productivity in the distant past. Groups who located in regions with uniformly low productivity might have done so due to a lifestyle of subsistence that does not rely on agriculture or inter-group exchange (e.g., hunter-gatherer societies). The distinction made here is important because regions with uniformly low land productivity are by definition low-variation regions, as is evident from the balancedness test in Figure A1. Although I control for pre-Columbian land productivity and the post-Columbian change in land productivity throughout the entire analysis, this only guarantees the trade mechanism holds conditional on the productivity mechanism, rather than ruling out it out as a competing explanation.

To test this competing mechanism, I first create three indicator variables denoting the following of a buffer zone: (i) the location is or is not uniformly low in productivity, (ii) the location is or is not uniformly high in productivity and (iii) the location is or is not a mixed productivity region. Uniformly low (high) productivity region types are defined by a two-step process. I define low (high) productivity regions as any region in the bottom (top) decile of land productivity. Then, conditional on being a low (high) productivity region, I define *uniformly* low (high) productivity regions as those below some threshold for variation in land productivity. In some applications, I set this threshold to include low (high) productivity regions that are also in the bottom decile of land productivity variation, and in others I set the threshold as those in the bottom quartile of land productivity variation. I define mixed productivity regions as buffer zones that are neither uniformly low nor uniformly high in productivity. All references to productivity and productivity variation here correspond to pre-1500 CE levels.

As a test of this competing mechanism I first compare mean differences in linguistic distance for groups in/out of a defined productivity region type. Table A9 reports these results. In all instances, across all definitions of region type, I find no significant differences in linguistic distance. This is the first piece of evidence that rules out low land productivity as a competing explanation.

As a second test, I consider what happens to the relationship between variation in land productivity and linguistic distance when regions with uniformly low land productivity are dropped from the full sample and sibling sample. Table A10 reports these estimates. Column 1 is a replication of column 6 in Table 1—my preferred baseline estimate. For the estimates in column 2, I drop *all* low productivity regions regardless of whether they are classified as *uniformly* low or not. For the estimates in columns 3 and 4, I drop only uniformly low productivity regions, based on the alternative thresholds for low land productivity variation. The estimates reported in column 5-8 reflect the analogous set of estimates for the sibling sample. In all instances, the coefficients of interest increase in magnitude while the standard errors remain relatively constant, resulting in even more precise estimates than at baseline. Overall, these findings rule out this alternative channel and give further credibility to the trade mechanism.

Aggregate Population and Differences—In a Malthusian world, variation in land productivity will correspond to variation in population. Throughout the empirical analysis I control for the aggregate population of the two groups associated with each buffer zone, but I do not consider how population dynamics influence my baseline result. For example, large populations may be less susceptible the influence of an outside group, or perhaps influence is weakened when adjacent groups are similar in population size. With this in mind, I make two distinctions based on the population associated with each buffer zone: (i) the aggregate population of both groups and (ii) the difference in population between groups. For each distinction, I test the sensitivity of the baseline result to dropping all buffer zones above the median.

Table A11 reports these estimates. All coefficients have the expected negative sign, and most actually increase in magnitude. The estimates in columns 2 and 5 suggest that buffer zones with below-median aggregate populations tend to be more influenced by the geographical-trade mechanism compared to the full sample of observations. Whereas the evidence is mixed with respect to population differences. For the full sample (column 3), the estimate is more than double the comparable baseline estimate, indicating that groups most affected by the trade mechanism tend to be similar in population size (i.e., below the median in population differences). The sibling-sample estimates are similarly larger than at baseline (column 6), but are estimated with less precision so they fall outside of a standard level of confidence, hence the evidence on this front is mixed.

Overlapping Polygons—In the *Ethnologue*, the homeland of ethnolinguistic groups sometimes overlap. This is problematic because any geographic variable specific to a buffer zone will not uniquely represent the associated group pair. In Table A12, for both samples, I compare the baseline estimate to a subsample estimate that excludes all overlapping border buffer zones. I find no evidence that the main empirical result is a consequent of overlapping polygons.

Selection on Unobservables—Table A13 reports Altonji et al. (2005) (AET) statistics on how much stronger selection on unobservables must be compared to selection on observables in order to fully explain the results. To perform this test, I calculate the ratio $\hat{\beta^F}/(\hat{\beta^R} - \hat{\beta^F})$, where $\hat{\beta^F}$ is the coefficient estimate of the variable of interest—Δ in land productivity variation (pre-1500)—that includes a full set of controls, while $\hat{\beta^R}$ is the estimated coefficient of the variable of interest for the restricted set of controls. Overall, the reported AET statistics suggest the influence of unobservable characteristics must be 1.35 to 4.26 times stronger than the influence of observables to fully account for my baseline findings.

As an additional check for selection on unobservables, I use a method developed by Oster (2019). Assuming that unobservables are as important in explaining the outcome variable as observables, Oster derives a bias-adjusted estimate for the coefficient of interest. I calculate this lower bound using the formula $\beta^* = \hat{\beta^F} - (\hat{\beta^R} - \hat{\beta^F}) \times \frac{R^2_{max} - R^2_F}{R^2_F - R^2_R}$, where $\hat{\beta^F}$ and $\hat{\beta^R}$ are defined the same as above, R^2_F is the R-squared from the fully-controlled regression and R^2_R is the R-squared from the restricted regression. I set $R^2_{max} = 1$, the theoretical maximum, even though the presence of measurement error in the dependent variable suggests this is an overly conservative assumption. I report β^* in Table A13. Even under the most conservative assumption of $R^2_{max} = 1$, in all instances the coefficient remains negative and sizable in magnitude.

Spatial Correlation—In Table A14, I report estimates of the baseline model with country fixed effects but allow for spatial correlation in the error term. I find that the coefficient of interest remains statistically significant at spatial correlation cutoffs of 100km, 200km, 300km, 400km and 500km. I use the Stata package `acreg` to make these standard error adjustments (Colella et al., 2019).

Placebo Tests—I also run a variety of placebo tests, where I limit the sample to geographic regions inhospitable to trade (i.e., sample observations above the 90th percentile in elevation or ruggedness), and New World observations where post-Columbian migrations introduce significant noise to the estimates. The hypothesized trade mechanism is expected to disappear in regions inhospitable to trade, hence breaking the link between land productivity variation and linguistic distance. Whereas any long-run evidence of the trade mechanism should be washed away by significant post-Columbian migration in the New World sample. Table A15 reports the results of these tests, where the baseline result disappears in both the full sample and sibling sample as expected.

Border-Level Regressions: Linguistic Distance and Trade Route Proximity

Here, I document suggestive evidence of the long-run impact of a group's exposure to interethnic trade on language. For this part of the empirical analysis, I rely on an empirical model similar to equation (4) in the main text. I again define buffer zone k as the region surrounding the segment of border that separates ethnolinguistic groups i and j. My main independent variables include measures of geodesic distance between buffer zone k's centroid and the nearest trade route for two time periods: pre-600 CE and pre-1800 CE.[1] This approach is motivated by the evidence that pre-colonial ethnic groups located near Old World trade routes were more likely to engage in trade (Michalopoulos et al., 2018). I estimate the relationship between historic trade route proximity and contemporary linguistic distance between groups i and j in the following way:

$$LD_k = \mu_0 + \mu Distance_k + x_k' \rho + \lambda_{l_i(k)} + \theta_{l_j(k)} + \delta_{c(k)} + \epsilon_k. \tag{1}$$

The dependent variable LD_k measures the linguistic distance between neighboring ethnolinguistic groups i and j in buffer zone k. Depending on the specification, $Distance_k$ captures the (logged) distance between buffer zone k and the nearest Old World trade route in the pre-600 CE period or the pre-1800 CE period. x_k represents a vector of buffer zone geo-climatic characteristics identical to the complete set of controls included in the baseline analysis, $\lambda_{l_i(k)}$

[1] I drop all observations associated with New World countries since I rely on Old World trade maps. Digitized trade route maps come from Michalopoulos et al. (2016, 2018).

and $\lambda_{l_j(k)}$ respectively denote language family fixed effects for group i and j, and $\delta_{c(k)}$ represents country fixed effects. If groups located near trade routes more actively engaged in trade, and inter-ethnic trade mitigates cultural drift, then it is expected that $\hat{\mu} > 0$.

Table A16 reports estimates of equation (1) for both the full sample and sibling sample. Both pre-600 CE distance and pre-1800 CE distance coefficients are positive as expected, but the positive association between pre-1800 trade route distance and linguistic distance is far more robust. Regardless of specification or sample, pre-1800 distance retains statistical significance at conventional levels (columns 3 and 4), and outperforms pre-600 distance in a horse race regression (column 5).

Yet for 25 percent of the full sample and 17 percent of the sibling sample observations, the distance to the nearest trade route is unchanged between 600 CE and 1800 CE. For these observations, both distance measures are statistically indistinguishable in a regression analysis. To get around this, I also calculate the change in log distance between 600 CE and 1800 CE, and re-estimate equation (1) using pre-600 distance in levels and the post-600 change in distance. These estimates are reported in column 6, where both measures are now positive and statistically significant. For each sample, the coefficient associated with the post-600 change in column 6 is identical to the pre-1800 coefficient in column 5, since these variables are estimated using identical variation. However, in the column 6 estimates, pre-600 observations with unchanged distance in the post-600 period no longer suffer from a collinearity problem, and are estimated to be positive and significant.

Overall, these data suggest that neighboring ethnic groups historically located near pre-600 and pre-1800 Old World trade routes are more similar in language today than neighboring groups located further away. This finding is consistent with evidence that groups living near historical trade routes engaged in trade more than those living further away (Michalopoulos et al., 2018). While these estimates do not rely on post-Columbian changes, and thus cannot necessarily be interpreted as causal, they are still informative of the channel through which historical trade shapes contemporary differences in language.

Table A1: Summary Statistics – Border-Level Dataset

	Mean	Std dev.	Min	Max	*N*
	Panel A: Full Sample				
Linguistic distance	0.727	0.177	0.003	0.945	8402
Land productivity variation pre-1500	0.236	0.272	0	1.616	8402
Δ in land productivity variation post-1500	-0.08	0.201	-0.977	0.374	8402
Land productivity pre-1500	1.41	0.752	0	5.151	8402
Δ in land productivity post-1500	-0.118	0.624	-2.136	1.388	8402
Malaria	8.314	8.891	0	36.286	8402
Elevation	691.059	655.778	-22.303	4929.123	8402
Ruggendess	306.502	287.681	0.101	1931.97	8402
Precipitation	14.202	7.935	0.011	50.669	8402
Precipitation variation	1.714	1.795	0	19.648	8402
Temperature	21.912	6.431	-12.287	29.492	8402
Temperature variation	1.591	1.481	0.037	10.092	8402
Log distance between group centroids	4.656	1.405	0	8.637	8402
Log distance to coast	4.928	1.839	-3.797	7.697	8402
Log distance to border	3.032	2.218	-12.096	6.838	8402
Log distance to lake	4.747	1.034	0	7.977	8402
Log distance to major river	4.259	1.707	-5.554	8.975	8402
Log distance to minor river	5.663	1.912	-5.147	8.976	8402
Log population of group pair	12.578	3.482	1.609	20.637	8402
Log area of group pair	9.707	2.705	2.18	15.996	8402
Difference in absolute latitude	1.435	2.523	0	30.337	8402
Difference in absolute longitude	2.127	6.94	0	340.317	8402

	Mean	Std dev.	Min	Max	*N*
	Panel B: Sibling Sample				
Linguistic distance	0.525	0.168	0.003	0.913	1669
Land productivity variation pre-1500	0.295	0.306	0	1.373	1669
Δ in land productivity variation post-1500	-0.149	0.263	-0.977	0.252	1669
Land productivity pre-1500	1.475	0.737	0	4.598	1669
Δ in land productivity post-1500	-0.245	0.73	-2.064	1.335	1669
Malaria	8.938	8.852	0	35.624	1669
Elevation	680.91	636.435	1.057	4929.123	1669
Ruggendess	336.121	295.268	0.295	1894.648	1669
Precipitation	16.12	8.194	0.276	48.676	1669
Precipitation variation	1.908	1.764	0.038	16.69	1669
Temperature	22.77	5.056	-6.977	29.21	1669
Temperature variation	1.72	1.514	0.037	9.981	1669
Log distance between group centroids	3.734	1.185	0	8.092	1669
Log distance to coast	4.401	1.945	-2.944	7.576	1669
Log distance to border	3.029	1.855	-9.778	6.581	1669
Log distance to lake	4.82	0.963	0	7.411	1669
Log distance to major river	4.747	1.722	-3.61	7.866	1669
Log distance to minor river	6.387	1.89	-3.61	8.82	1669
Log population of group pair	10.55	2.871	2.303	20.637	1669
Log area of group pair	7.993	2.202	2.652	15.971	1669
Difference in absolute latitude	0.511	0.997	0	15.229	1669
Difference in absolute longitude	0.594	2.001	0	51.597	1669

Table A2: Summary Statistics – Ancestral Characteristics Group-Level Dataset

	Mean	Std dev.	Min	Max	*N*
=1 if agriculture is dominant subsistence activity	0.824	0.381	0	1	6128
=1 if pastoralism is dominant subsistence activity	0.046	0.21	0	1	6128
=1 if fishing is dominant subsistence activity	0.038	0.192	0	1	6128
=1 if hunting-gathering is dominant subsistence activity	0.092	0.288	0	1	6128
=1 if inter-ethnic trade is important for subsistence	0.897	0.304	0	1	688
=1 if rely on inter-ethnic trade as food source	0.724	0.447	0	1	1096
=1 if exogamous community (EA)	0.069	0.254	0	1	5458
=1 if exogamous community (SCCS)	0.167	0.373	0	1	1161
=1 if frequently attacking other groups	0.747	0.435	0	1	899
=1 if frequenly being attached by other groups	0.604	0.489	0	1	869
Land productivity pre-1500	1.365	0.784	0	4.975	6128
∆ in land productivity post-1500	-0.101	0.674	-2.277	1.536	6128
Land productivity variation pre-1500	0.131	0.216	0	2.048	6128
∆ in land productivity variation post-1500	-0.033	0.147	-1.253	0.879	6128
Elevation	683.833	770.111	-4.925	5631.069	6128
Ruggendess	198.33	215.576	0	1710.952	6128
Precipitation	14.219	8.127	0	56.433	6128
Precipitation variation	0.982	1.267	0	19.985	6128
Temperature	21.979	6.784	-14.784	29.723	6128
Temperature variation	0.999	1.119	0	8.710	6128
Log distance to coast	11.708	1.861	3.326	14.597	6128
Log distance to border	10.468	1.526	2.806	13.751	6128
Log distance to lake	11.66	1.313	0	15.042	6128
Log distance to major river	11.443	1.584	1.279	15.886	6128
Log distance to minor river	12.765	1.747	1.279	15.886	6128
Log population	9.128	3.153	0	20.549	6128
Log land area	7.115	1.919	2.717	15.899	6128
Latitude of group centroid	8.332	17.295	-40.29	73.273	6128
Longitude of group centroid	46.884	77.588	-178.787	179.311	6128

Table A3: Border-Level Regressions: Full Sample Baseline Results (100km Buffer Zone)

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$

	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	-0.100***	-0.100***	-0.083**	-0.103***	-0.098***	-0.065*
	(0.030)	(0.034)	(0.033)	(0.033)	(0.033)	(0.034)
Land productivity variation (pre-1500)	-0.061***	-0.061***	-0.046*	-0.043**	-0.040	-0.029
	(0.022)	(0.022)	(0.026)	(0.021)	(0.026)	(0.027)
Δ in land productivity (post-1500)		-0.001	0.001	-0.005	-0.001	0.008
		(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Land productivity (pre-1500)		-0.001	-0.001	0.001	0.004	0.009
		(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Malaria Ecology Index			0.003***		0.004***	0.001**
			(0.001)		(0.001)	(0.001)
Elevation			-0.000		0.000	0.000
			(0.000)		(0.000)	(0.000)
Ruggedness			0.000**		0.000***	0.000***
			(0.000)		(0.000)	(0.000)
Average precipitation			-0.000		0.000	-0.001
			(0.001)		(0.001)	(0.001)
Precipitation variation			-0.003		-0.004**	-0.004*
			(0.002)		(0.002)	(0.002)
Average temperature			0.000		0.000	0.001
			(0.002)		(0.002)	(0.002)
Temperature variation			-0.016*		-0.015*	-0.018**
			(0.008)		(0.008)	(0.008)
Log distance between group centroids				0.024***	0.025***	0.026***
				(0.005)	(0.004)	(0.005)
Log distance to coast				-0.004	-0.007**	0.005*
				(0.003)	(0.003)	(0.003)
Log distance to border				-0.000	-0.001	-0.008***
				(0.001)	(0.001)	(0.002)
Log distance to lake				0.006**	0.006*	0.005
				(0.003)	(0.003)	(0.003)
Log distance to major river				0.006***	0.006***	0.001
				(0.002)	(0.002)	(0.002)
Log distance to minor river				-0.008***	-0.005*	-0.001
				(0.003)	(0.003)	(0.003)
Log total population				0.002	0.003	0.000
				(0.002)	(0.002)	(0.002)
Log total area				-0.003	-0.002	0.009**
				(0.004)	(0.003)	(0.004)
Latitude difference				-0.001	-0.002	-0.003**
				(0.001)	(0.001)	(0.001)
Longitude difference				0.000	0.000	-0.000
				(0.000)	(0.000)	(0.000)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	No	No	Yes
Adjusted R^2	0.25	0.25	0.26	0.26	0.28	0.37
Observations	8402	8402	8402	8402	8402	7291

Unit of observation: border buffer zone (100km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair's lexicostatistical linguistic distance. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Border-Level Regressions: Sibling Sample Baseline Results (100km Buffer Zone)

	Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$					
	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	-0.188*** (0.042)	-0.197*** (0.047)	-0.273*** (0.060)	-0.185*** (0.047)	-0.248*** (0.060)	-0.154** (0.070)
Land productivity variation (pre-1500)	-0.088** (0.034)	-0.088** (0.034)	-0.191*** (0.059)	-0.097*** (0.035)	-0.185*** (0.059)	-0.140** (0.068)
Δ in land productivity (post-1500)		-0.007 (0.016)	-0.009 (0.015)	-0.001 (0.016)	-0.003 (0.015)	0.010 (0.019)
Land productivity (pre-1500)		-0.018 (0.012)	-0.008 (0.011)	-0.018 (0.012)	-0.009 (0.012)	0.011 (0.014)
Malaria Ecology Index			0.002** (0.001)		0.002* (0.001)	-0.001 (0.001)
Elevation			0.000 (0.000)		0.000 (0.000)	0.000 (0.000)
Ruggedness			0.000* (0.000)		0.000* (0.000)	0.000 (0.000)
Average precipitation			-0.001 (0.001)		-0.001 (0.001)	-0.003** (0.001)
Precipitation variation			-0.003 (0.005)		-0.000 (0.005)	-0.002 (0.005)
Average temperature			0.003 (0.003)		0.002 (0.003)	0.007 (0.006)
Temperature variation			-0.016 (0.017)		-0.013 (0.017)	-0.008 (0.018)
Log distance between group centroids				0.024** (0.009)	0.023** (0.009)	0.017* (0.009)
Log distance to coast				0.007 (0.005)	0.005 (0.005)	0.008 (0.007)
Log distance to border				0.000 (0.003)	-0.001 (0.003)	-0.007 (0.006)
Log distance to lake				0.015*** (0.005)	0.014** (0.006)	0.011* (0.006)
Log distance to major river				0.005 (0.004)	0.004 (0.004)	-0.002 (0.005)
Log distance to minor river				-0.001 (0.006)	0.001 (0.006)	0.015* (0.008)
Log total population				-0.004 (0.004)	-0.005 (0.004)	-0.006 (0.004)
Log total area				-0.019*** (0.006)	-0.017*** (0.006)	0.002 (0.007)
Latitude difference				0.001 (0.005)	0.002 (0.005)	-0.005 (0.005)
Longitude difference				0.002* (0.001)	0.003** (0.001)	0.008 (0.006)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	No	No	Yes
Adjusted R^2	0.28	0.28	0.29	0.30	0.30	0.39
Observations	1669	1669	1669	1669	1669	1497

Unit of observation: border buffer zone (100km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair's lexicostatistical linguistic distance for the baseline sibling sample. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Group-Level Regressions: Land Productivity and Historical Modes of Subsistence

	Dependent Variables: Dominant Historical Subsistence Activity			
	Agriculture	Pastoralism	Fishing	Hunter-Gatherer
	(1)	(2)	(3)	(4)
Δ in land productivity variation (post-1500)	-0.085	0.028	-0.005	0.062
	(0.080)	(0.066)	(0.029)	(0.038)
Land productivity variation (pre-1500)	-0.080	0.010	0.004	0.066
	(0.081)	(0.063)	(0.023)	(0.046)
Δ in land productivity (post-1500)	0.092***	-0.042*	-0.027*	-0.023
	(0.031)	(0.024)	(0.014)	(0.020)
Land productivity (pre-1500)	0.107***	-0.045***	-0.034***	-0.027
	(0.028)	(0.017)	(0.013)	(0.023)
Elevation	0.000	0.000	-0.000***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Ruggedness	-0.000	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Average precipitation	-0.001	-0.002**	-0.000	0.003*
	(0.003)	(0.001)	(0.001)	(0.002)
Precipitation variation	0.009*	-0.002	-0.000	-0.006
	(0.005)	(0.003)	(0.005)	(0.004)
Average temperature	-0.000	0.005***	-0.011**	0.006
	(0.004)	(0.002)	(0.005)	(0.006)
Temperature variation	0.003	0.012	-0.009	-0.006
	(0.020)	(0.012)	(0.008)	(0.018)
Log distance to coast	0.010	-0.001	-0.007	-0.003
	(0.009)	(0.004)	(0.006)	(0.009)
Log distance to border	0.013*	-0.008*	-0.006	0.001
	(0.007)	(0.004)	(0.004)	(0.004)
Log distance to lake	0.010*	-0.003	-0.007*	-0.000
	(0.005)	(0.003)	(0.004)	(0.003)
Log distance to major river	0.008	-0.001	-0.005	-0.002
	(0.005)	(0.003)	(0.004)	(0.005)
Log distance to minor river	-0.010	0.009	-0.008	0.009*
	(0.008)	(0.005)	(0.007)	(0.005)
Log total population	0.012***	-0.003**	-0.002	-0.007***
	(0.003)	(0.001)	(0.002)	(0.002)
Log total area	-0.020***	0.012***	0.001	0.006
	(0.005)	(0.004)	(0.002)	(0.004)
Latitude	-0.001	-0.001	-0.000	0.002
	(0.003)	(0.001)	(0.003)	(0.003)
Longitude	-0.001*	-0.000	0.000	0.001
	(0.001)	(0.000)	(0.001)	(0.001)
Language Family FE	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.54	0.37	0.28	0.62
Observations	6128	6128	6128	6128

Unit of observation: ethnolinguistic group. This table illustrates the effect of land productivity on historical modes of subsistence. Each dependent variable is an indicator that takes a value of one if that mode of subsistence is a group's dominant historical subsistence activity. Robust standard errors are clustered at the country level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Group-Level Regressions: Land Productivity Variations and Social Interactions

	Dependent Variables: Social Interactions					
	Inter-Ethnic Trade		Inter-Ethnic Marriage		Inter-Ethnic Conflict	
	For Subsistence	For Food	Exogamy (EA Sample)	Exogamy (SCCS Sample)	Frequently Attacks Others	Frequently Is Attacked
	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	0.201**	0.316**	0.114**	0.375*	-0.502**	-0.578**
	(0.097)	(0.157)	(0.056)	(0.199)	(0.218)	(0.232)
Land productivity variation (pre-1500)	0.139*	0.165	0.072	0.260	-0.349*	-0.472**
	(0.079)	(0.105)	(0.052)	(0.165)	(0.183)	(0.216)
Δ in land productivity (post-1500)	0.022	0.128	-0.005	-0.167**	0.005	0.045
	(0.025)	(0.084)	(0.020)	(0.083)	(0.068)	(0.069)
Land productivity (pre-1500)	-0.002	0.063	-0.038*	-0.113**	-0.037	0.006
	(0.019)	(0.046)	(0.021)	(0.052)	(0.045)	(0.052)
Elevation	-0.000	0.000	-0.000	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Ruggedness	-0.000	-0.000	-0.000	0.000	-0.000*	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average precipitation	-0.002	0.003	-0.004*	0.000	0.007**	0.008**
	(0.002)	(0.003)	(0.002)	(0.004)	(0.004)	(0.003)
Precipitation variation	-0.002	0.007	-0.001	0.007	-0.007	0.006
	(0.003)	(0.013)	(0.005)	(0.012)	(0.016)	(0.014)
Average temperature	-0.003	0.005	0.005	-0.003	0.017**	0.005
	(0.005)	(0.013)	(0.004)	(0.008)	(0.009)	(0.004)
Temperature variation	0.014	0.018	0.003	-0.008	0.066**	-0.023
	(0.010)	(0.022)	(0.012)	(0.034)	(0.032)	(0.034)
Log distance to coast	0.010	-0.022	0.007	-0.016	0.081***	0.039**
	(0.016)	(0.019)	(0.007)	(0.017)	(0.023)	(0.016)
Log distance to border	0.009	0.001	-0.012*	0.022	-0.028*	0.004
	(0.012)	(0.016)	(0.007)	(0.017)	(0.015)	(0.011)
Log distance to lake	-0.009**	0.008	0.005	0.011	-0.021*	0.012
	(0.004)	(0.009)	(0.005)	(0.011)	(0.012)	(0.010)
Log distance to major river	0.010	0.029	-0.003	-0.018	0.042*	0.010
	(0.012)	(0.021)	(0.008)	(0.019)	(0.022)	(0.017)
Log distance to minor river	-0.014	-0.026	0.020***	0.010	-0.024	-0.032
	(0.014)	(0.022)	(0.007)	(0.020)	(0.023)	(0.020)
Log total population	0.004	0.005	-0.000	-0.008	0.006	0.004
	(0.004)	(0.004)	(0.002)	(0.005)	(0.006)	(0.005)
Log total area	-0.008**	-0.009	0.001	0.002	0.005	0.005
	(0.004)	(0.008)	(0.004)	(0.009)	(0.010)	(0.010)
Latitude	-0.006	-0.002	0.000	-0.005	0.002	-0.000
	(0.005)	(0.006)	(0.002)	(0.006)	(0.003)	(0.005)
Longitude	-0.004*	-0.000	-0.000	0.003*	0.001	0.008***
	(0.002)	(0.002)	(0.001)	(0.002)	(0.004)	(0.003)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.90	0.75	0.29	0.68	0.66	0.78
Observations	688	1096	5458	1161	899	869

Unit of observation: ethnolinguistic group. This table illustrates the effect of land productivity variations on inter-ethnic trade and other forms of social interactions. Each dependent variable is an indicator that takes a value of one for group's who engage in the defined inter-ethnic social interaction. Robust standard errors are clustered at the country level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Border-Level Regressions: Full Sample Baseline Results (50km Buffer Zone)

	Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$					
	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	-0.115*** (0.033)	-0.107*** (0.034)	-0.107*** (0.034)	-0.118*** (0.034)	-0.118*** (0.034)	-0.058* (0.032)
Land productivity variation (pre-1500)	-0.090*** (0.024)	-0.089*** (0.024)	-0.090*** (0.024)	-0.076*** (0.023)	-0.077*** (0.023)	-0.033 (0.023)
Δ in land productivity (post-1500)		-0.011 (0.010)	-0.011 (0.010)	-0.016 (0.010)	-0.016 (0.010)	0.003 (0.013)
Land productivity (pre-1500)		-0.006 (0.007)	-0.006 (0.007)	-0.003 (0.007)	-0.003 (0.007)	0.012 (0.008)
Malaria Ecology Index			0.000 (0.000)		0.000 (0.000)	-0.000 (0.000)
Elevation			0.000 (0.000)		0.000* (0.000)	0.000 (0.000)
Ruggedness			-0.000** (0.000)		-0.000** (0.000)	-0.000 (0.000)
Average precipitation			0.000 (0.000)		0.001* (0.000)	0.000 (0.000)
Precipitation variation			0.000 (0.002)		0.000 (0.002)	0.001 (0.002)
Average temperature			0.000 (0.001)		0.000 (0.001)	0.000 (0.001)
Temperature variation			0.015** (0.006)		0.013** (0.006)	0.008 (0.006)
Log distance between group centroids				0.024*** (0.005)	0.024*** (0.005)	0.026*** (0.005)
Log distance to coast				-0.004 (0.003)	-0.004 (0.003)	0.008*** (0.003)
Log distance to border				0.000 (0.001)	0.000 (0.001)	-0.009*** (0.002)
Log distance to lake				0.006** (0.003)	0.006** (0.003)	0.006** (0.003)
Log distance to major river				0.005** (0.002)	0.005** (0.002)	0.000 (0.002)
Log distance to minor river				-0.008*** (0.003)	-0.008*** (0.003)	-0.003 (0.003)
Log total population				0.002 (0.002)	0.002 (0.002)	-0.001 (0.002)
Log total area				-0.003 (0.004)	-0.003 (0.004)	0.009*** (0.004)
Latitude difference				-0.001 (0.001)	-0.001 (0.001)	-0.003** (0.001)
Longitude difference				0.000 (0.000)	0.000 (0.000)	-0.000 (0.000)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	No	No	Yes
Adjusted R^2	0.25	0.25	0.25	0.26	0.27	0.37
Observations	8402	8402	8402	8402	8402	7290

Unit of observation: border buffer zone (50km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair's lexicostatistical linguistic distance. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8: Border-Level Regressions: Post-Columbian Migrations Sensitivity Analysis

	Dependent Variable: Lexicostatistical Linguistic Distance $\in (0,1)$							
	Full Sample				Sibling Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ in land productivity variation (post-1500)	-0.065*	-0.072**	-0.073**	-0.078**	-0.154**	-0.159**	-0.154**	-0.175**
	(0.034)	(0.036)	(0.036)	(0.038)	(0.070)	(0.071)	(0.071)	(0.073)
Land productivity variation (pre-1500)	-0.029	-0.032	-0.031	-0.059*	-0.140**	-0.142**	-0.138**	-0.179**
	(0.027)	(0.029)	(0.030)	(0.032)	(0.068)	(0.069)	(0.069)	(0.072)
Δ in land productivity (post-1500)	0.008	0.010	0.010	0.006	0.010	0.013	0.012	0.004
	(0.010)	(0.012)	(0.012)	(0.013)	(0.019)	(0.020)	(0.021)	(0.022)
Land productivity (pre-1500)	0.009	0.012*	0.014**	0.017**	0.011	0.012	0.012	0.014
	(0.006)	(0.007)	(0.007)	(0.008)	(0.014)	(0.015)	(0.016)	(0.017)
Malaria Ecology Index	0.001**	0.001*	0.001**	0.001	-0.001	-0.002	-0.001	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Elevation	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Ruggedness	0.000***	0.000***	0.000**	0.000**	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average precipitation	-0.001	-0.001	-0.001	-0.001	-0.003**	-0.003**	-0.003**	-0.003**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Precipitation variation	-0.004*	-0.004*	-0.004	-0.004	-0.002	-0.001	-0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.003)	(0.005)	(0.005)	(0.005)	(0.005)
Average temperature	0.001	0.001	0.001	0.001	0.007	0.009	0.010	0.011
	(0.002)	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.007)
Temperature variation	-0.018**	-0.019**	-0.017*	-0.024**	-0.008	-0.009	-0.010	-0.002
	(0.008)	(0.009)	(0.009)	(0.012)	(0.018)	(0.018)	(0.018)	(0.023)
Log distance between group centroids	0.026***	0.025***	0.025***	0.024***	0.017*	0.018*	0.019**	0.020*
	(0.005)	(0.005)	(0.005)	(0.005)	(0.009)	(0.009)	(0.009)	(0.010)
Log distance to coast	0.005*	0.005	0.005	0.005	0.008	0.007	0.007	0.008
	(0.003)	(0.003)	(0.003)	(0.004)	(0.007)	(0.007)	(0.007)	(0.007)
Log distance to border	-0.008***	-0.009***	-0.009***	-0.010***	-0.007	-0.007	-0.007	-0.007
	(0.002)	(0.002)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.006)
Log distance to lake	0.005	0.005	0.006*	0.006*	0.011*	0.011*	0.011*	0.012*
	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)	(0.006)	(0.006)	(0.007)
Log distance to major river	0.001	0.000	0.001	0.000	-0.002	-0.002	-0.001	-0.002
	(0.002)	(0.002)	(0.003)	(0.003)	(0.005)	(0.005)	(0.005)	(0.006)
Log distance to minor river	-0.001	-0.001	-0.002	-0.001	0.015*	0.016*	0.014*	0.011
	(0.003)	(0.003)	(0.003)	(0.004)	(0.008)	(0.008)	(0.008)	(0.008)
Log total population	0.000	0.001	0.001	0.000	-0.006	-0.006	-0.005	-0.007
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)
Log total area	0.009**	0.009**	0.009**	0.010**	0.002	-0.000	-0.001	0.002
	(0.004)	(0.004)	(0.004)	(0.004)	(0.007)	(0.007)	(0.007)	(0.007)
Latitude difference	-0.003**	-0.002	-0.002	-0.001	-0.005	-0.004	-0.004	-0.004
	(0.001)	(0.002)	(0.002)	(0.002)	(0.005)	(0.005)	(0.005)	(0.005)
Longitude difference	-0.000	0.000	0.000	0.001	0.008	0.008	0.007	0.007
	(0.000)	(0.001)	(0.001)	(0.001)	(0.006)	(0.006)	(0.006)	(0.007)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.37	0.36	0.36	0.34	0.39	0.40	0.40	0.39
Observations	7291	6723	6425	5540	1497	1445	1424	1273
Sample Restrictions								
Excluding regions with $\leq$ 25% native popluation	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Excluding regions with $\leq$ 50% native popluation	No	No	Yes	Yes	No	No	Yes	Yes
Excluding regions with $\leq$ 75% native popluation	No	No	No	Yes	No	No	No	Yes

Unit of observation: border buffer zone (100km). This table establishes the robustness of the baseline estimates to post-Columbian migrations with the full sample (columns 1-4) and sibling sample (columns 5-8). Each sample is restricted to language pairs that reside in a country where 25, 50 or 75 percent of the population is native to the residing country. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9: Difference in Means: Linguistic Differences By Productivity Region Type

	Panel A: Low Variation Threshold = 10th Percentile		
	Uniformly Low Productivity Region	Uniformly High Productivity Region	Mixed Productivity Region
$\mathbb{I}$(region type) $= 1$	0.731	0.709	0.727
$\mathbb{I}$(region type) $= 0$	0.727	0.727	0.731
Difference	0.004	-0.018	-0.004
p-value	0.701	0.860	0.716
	Panel B: Low Variation Threshold = 25th Percentile		
	Uniformly Low Productivity Region	Uniformly High Productivity Region	Mixed Productivity Region
$\mathbb{I}$(region type) $= 1$	0.735	0.688	0.727
$\mathbb{I}$(region type) $= 0$	0.727	0.727	0.733
Difference	0.008	-0.039	-0.006
p-value	0.361	0.309	0.515

Unit of observation: border buffer zone (100km). This table reports mean differences in linguistic distance for various productivity region types using the baseline full sample. Uniformly low (high) productivity region types are defined by a two-step process. I first define low (high) productivity regions as any region in the bottom (top) decile of land productivity. Then, conditional on being a low (high) productivity region, I define *uniformly* low (high) productivity regions as those below some threshold of variation in land productivity. In Panel A, I define uniformly low (high) productivity regions as any region below the 10th percentile of the distribution in productivity variation, conditional on being low (high) productivity. In Panel B, I do the same but set the threshold for low land productivity variation at the 25th percentile. In both panels, mixed productivity regions are any region that doesn't fall within the uniformly low or high productivity region classification. All productivity data corresponds to pre-1500 CE levels.

Table A10: Border-Level Regressions: Low Productivity Sensitivity Analysis

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$

	Full Sample				Sibling Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ in land productivity variation (post-1500)	-0.065*	-0.079**	-0.076**	-0.074**	-0.154**	-0.156**	-0.158**	-0.158**
	(0.034)	(0.040)	(0.035)	(0.036)	(0.070)	(0.076)	(0.070)	(0.071)
Land productivity variation (pre-1500)	-0.029	-0.048	-0.048	-0.044	-0.140**	-0.124*	-0.147**	-0.143**
	(0.027)	(0.034)	(0.030)	(0.030)	(0.068)	(0.075)	(0.069)	(0.069)
Δ in land productivity (post-1500)	0.008	0.003	0.002	0.007	0.010	0.029	0.008	0.016
	(0.010)	(0.012)	(0.011)	(0.011)	(0.019)	(0.022)	(0.020)	(0.021)
Land productivity (pre-1500)	0.009	0.004	0.006	0.008	0.011	0.023	0.013	0.016
	(0.006)	(0.007)	(0.007)	(0.007)	(0.014)	(0.017)	(0.016)	(0.016)
Malaria Ecology Index	0.001**	0.002**	0.001**	0.001**	-0.001	-0.002	-0.002	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Elevation	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Ruggedness	0.000***	0.000**	0.000***	0.000***	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average precipitation	-0.001	-0.002*	-0.001	-0.001	-0.003**	-0.003**	-0.003**	-0.003**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Precipitation variation	-0.004*	-0.003	-0.004*	-0.005*	-0.002	0.003	-0.002	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.005)	(0.005)	(0.005)	(0.005)
Average temperature	0.001	0.003	0.001	0.001	0.007	0.008	0.006	0.007
	(0.002)	(0.002)	(0.002)	(0.002)	(0.006)	(0.007)	(0.006)	(0.006)
Temperature variation	-0.018**	-0.016*	-0.021**	-0.021**	-0.008	-0.007	-0.012	-0.013
	(0.008)	(0.009)	(0.009)	(0.009)	(0.018)	(0.019)	(0.018)	(0.018)
Log distance between group centroids	0.026***	0.027***	0.028***	0.028***	0.017*	0.014	0.019*	0.020*
	(0.005)	(0.005)	(0.005)	(0.005)	(0.009)	(0.010)	(0.010)	(0.010)
Log distance to coast	0.005*	0.004	0.005	0.004	0.008	0.006	0.008	0.007
	(0.003)	(0.003)	(0.003)	(0.003)	(0.007)	(0.007)	(0.007)	(0.007)
Log distance to border	-0.008***	-0.009***	-0.008***	-0.008***	-0.007	-0.007	-0.006	-0.006
	(0.002)	(0.002)	(0.002)	(0.002)	(0.006)	(0.006)	(0.006)	(0.006)
Log distance to lake	0.005	0.005	0.004	0.005	0.011*	0.012*	0.010*	0.010
	(0.003)	(0.003)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)	(0.006)
Log distance to major river	0.001	0.000	0.001	0.001	-0.002	-0.004	-0.002	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.005)	(0.005)	(0.005)	(0.005)
Log distance to minor river	-0.001	-0.000	-0.001	-0.001	0.015*	0.013	0.014*	0.014*
	(0.003)	(0.003)	(0.003)	(0.003)	(0.008)	(0.009)	(0.008)	(0.008)
Log total population	0.000	0.001	0.000	0.001	-0.006	-0.006	-0.006	-0.006
	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)	(0.004)	(0.004)
Log total area	0.009**	0.008**	0.008**	0.007*	0.002	0.001	0.002	0.000
	(0.004)	(0.004)	(0.004)	(0.004)	(0.007)	(0.007)	(0.007)	(0.007)
Latitude difference	-0.003**	-0.002	-0.002**	-0.002**	-0.005	-0.006	-0.009	-0.009
	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)	(0.007)	(0.006)	(0.007)
Longitude difference	-0.000	-0.001	-0.001	-0.001	0.008	0.014	0.009	0.008
	(0.000)	(0.001)	(0.001)	(0.001)	(0.006)	(0.009)	(0.008)	(0.008)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.37	0.38	0.37	0.37	0.39	0.40	0.39	0.39
Observations	7291	6620	7096	7020	1497	1380	1469	1451
Sample Restrictions								
Excluding bottom decile land productivity	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Excluding bottom decile land productivity variation	No	No	Yes	No	No	No	Yes	No
Excluding bottom quartile land productivity variation	No	No	No	Yes	No	No	No	Yes

Unit of observation: border buffer zone (100km). This table establishes the robustness of the baseline estimates to dropping border buffer zones above the median in total population (columns 2 and 5) and buffer zones above the median in population differences (columns 3 and 6). Note columns 1 and 4 are reproductions of the baseline estimates from Table 1 and 2. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A11: Border-Level Regressions: Population Sensitivity Analysis

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$						
	Full Sample			Sibling Sample		
	(1)	(2)	(3)	(4)	(5)	(6)
Δ in land productivity variation (post-1500)	-0.065*	-0.192***	-0.177***	-0.154**	-0.290**	-0.186
	(0.034)	(0.056)	(0.052)	(0.070)	(0.133)	(0.156)
Land productivity variation (pre-1500)	-0.029	-0.183***	-0.165***	-0.140**	-0.350**	-0.236
	(0.027)	(0.055)	(0.050)	(0.068)	(0.135)	(0.158)
Δ in land productivity (post-1500)	0.008	0.003	0.008	0.010	0.015	0.005
	(0.010)	(0.016)	(0.015)	(0.019)	(0.027)	(0.028)
Land productivity (pre-1500)	0.009	0.004	0.010	0.011	0.047*	0.042
	(0.006)	(0.014)	(0.013)	(0.014)	(0.027)	(0.029)
Malaria Ecology Index	0.001**	0.001	0.001	-0.001	-0.003*	-0.003*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Elevation	0.000	-0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Ruggedness	0.000***	0.000**	0.000**	0.000	0.000*	0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Average precipitation	-0.001	-0.002*	-0.002**	-0.003**	-0.004**	-0.004**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
Precipitation variation	-0.004*	-0.004	-0.002	-0.002	-0.002	0.000
	(0.002)	(0.004)	(0.004)	(0.005)	(0.008)	(0.007)
Average temperature	0.001	-0.004	-0.001	0.007	0.007	0.013
	(0.002)	(0.005)	(0.004)	(0.006)	(0.012)	(0.013)
Temperature variation	-0.018**	-0.026	-0.025	-0.008	-0.040	-0.065*
	(0.008)	(0.016)	(0.016)	(0.018)	(0.034)	(0.033)
Log distance between group centroids	0.026***	0.040***	0.046***	0.017*	0.016	0.014
	(0.005)	(0.007)	(0.007)	(0.009)	(0.017)	(0.018)
Log distance to coast	0.005*	0.006	0.003	0.008	0.022**	0.017*
	(0.003)	(0.005)	(0.004)	(0.007)	(0.011)	(0.010)
Log distance to border	-0.008***	-0.010***	-0.008**	-0.007	-0.012	-0.006
	(0.002)	(0.003)	(0.003)	(0.006)	(0.010)	(0.009)
Log distance to lake	0.005	0.008*	0.003	0.011*	0.006	0.005
	(0.003)	(0.005)	(0.005)	(0.006)	(0.011)	(0.011)
Log distance to major river	0.001	-0.004	-0.001	-0.002	-0.002	-0.010
	(0.002)	(0.004)	(0.003)	(0.005)	(0.009)	(0.009)
Log distance to minor river	-0.001	0.006	0.008	0.015*	0.057**	0.038**
	(0.003)	(0.006)	(0.005)	(0.008)	(0.023)	(0.019)
Log total population	0.000	-0.006*	-0.010***	-0.006	0.002	-0.001
	(0.002)	(0.003)	(0.003)	(0.004)	(0.007)	(0.006)
Log total area	0.009**	0.008	0.006	0.002	0.011	0.013
	(0.004)	(0.005)	(0.005)	(0.007)	(0.010)	(0.011)
Latitude difference	-0.003**	-0.011	-0.014**	-0.005	-0.020	-0.025
	(0.001)	(0.007)	(0.007)	(0.005)	(0.028)	(0.026)
Longitude difference	-0.000	0.001	0.001	0.008	-0.008	-0.007
	(0.000)	(0.002)	(0.002)	(0.006)	(0.022)	(0.021)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	No	No	No	No	No	Yes
Adjusted R^2	0.40	0.37	0.38	0.44	0.42	0.41
Observations	7291	3616	3623	1497	733	729
Sample Restrictions						
Excluding buffer zones above median total population	No	Yes	No	No	Yes	No
Excluding buffer zones above median difference in population	No	No	Yes	No	No	Yes

Unit of observation: border buffer zone (100km). This table establishes the robustness of the baseline estimates to dropping border buffer zones above the median in total population (columns 2 and 5) and buffer zones above the median in population differences (columns 3 and 6). Note columns 1 and 4 are reproductions of the baseline estimates from Table 1 and 2. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12: Border-Level Regressions: Overlapping Polygon Sensitivity Analysis

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$				
	Full Sample		Sibling Sample	
	(1)	(2)	(3)	(4)
Δ in land productivity variation (post-1500)	-0.065*	-0.138***	-0.154**	-0.188**
	(0.034)	(0.043)	(0.070)	(0.081)
Land productivity variation (pre-1500)	-0.029	-0.095**	-0.140**	-0.186**
	(0.027)	(0.039)	(0.068)	(0.080)
Controls	Yes	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.40	0.40	0.44	0.42
Observations	7291	5099	1497	1262
Overlapping Polygons Excluded	No	Yes	No	Yes

Unit of observation: border buffer zone (100km). This table tests the sensitivity of the baseline estimates by limiting the sibling sample to ethnolinguistic pairs that do not overlap with any other groups. Control variables are identical to the complete set of baseline control variables used in Table 1. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A13: Border-Level Regressions: Selection on Unobservables

Controls in Restricted Set	Controls in Full Set	Altonji et al. (2005)	Oster (2019) β^*
Panel A: Full Sample			
FE	FE, Prod, Geog, Spatial	-3.69	-0.38
FE, Prod	FE, Prod, Geog, Spatial	-4.26	-0.34
Panel B: Sibling Sample			
FE	FE, Prod, Geog, Spatial	-1.35	-2.77
FE, Prod	FE, Prod, Geog, Spatial	-1.51	-2.64

Unit of observation: border buffer zone (100km). This table reports Altonji et al's (2005) measure of selection on unobservables and Oster's (2019) β^* lower bound estimate of the coefficient for the variable of interest: Δ in land productivity variation (post-1500). The dependent variable in each regression is lexicostatistical linguistic distance. FE = language family and country fixed effects, prod = productivity controls, geog = geography controls and spatial = spatial controls. These variables are equivalent to the baseline set described in Table 1.

Table A14: Border-Level Regressions: Spatially Correlated Standard Errors

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$					
	(1)	(2)	(3)	(4)	(5)
	Panel A: Full Sample				
Δ in land productivity variation (post-1500)	-0.065*	-0.065*	-0.065*	-0.065**	-0.065*
	(0.035)	(0.039)	(0.036)	(0.030)	(0.034)
Land productivity variation (pre-1500)	-0.029	-0.029	-0.029	-0.029	-0.029
	(0.028)	(0.030)	(0.028)	(0.025)	(0.030)
Centered R^2	0.40	0.40	0.40	0.40	0.40
Observations	7291	7291	7291	7291	7291
Controls	Yes	Yes	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes
Spatial Correlation Cutoff	100km	200km	300km	400km	500km
	Panel B: Sibling Sample				
Δ in land productivity variation (post-1500)	-0.154**	-0.154**	-0.154**	-0.154**	-0.154**
	(0.073)	(0.075)	(0.075)	(0.071)	(0.072)
Land productivity variation (pre-1500)	-0.140**	-0.140**	-0.140**	-0.140**	-0.140**
	(0.068)	(0.071)	(0.071)	(0.071)	(0.067)
Centered R^2	0.46	0.46	0.46	0.46	0.46
Observations	1497	1497	1497	1497	1497
Controls	Yes	Yes	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes
Spatial Correlation Cutoff	100km	200km	300km	400km	500km

Unit of observation: border buffer zone (100km). This table establishes the negative and statistically significant effect of variation in land productivity on a language pair's lexicostatistical linguistic distance is robust to adjusting for spatial correlation at various distance thresholds, across both the full sample and sibling sample. Control variables are identical to the baseline set of control variables used in Table 1. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A15: Border-Level Regressions: Placebo Tests

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0,1)$				
	(1)	(2)	(3)	(4)
	Panel A: Full Sample			
Δ in land productivity variation (post-1500)	-0.065*	-0.009	-0.044	-0.019
	(0.034)	(0.092)	(0.078)	(0.092)
Land productivity variation (pre-1500)	-0.029	0.054	0.044	0.058
	(0.027)	(0.044)	(0.046)	(0.068)
Controls	Yes	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.37	0.44	0.40	0.51
Observations	7291	708	711	1084
Sample Restriction	None	$\geq$ 90% Elevation	$\geq$ 90% Ruggedness	New World
	Panel B: Sibling Sample			
Δ in land productivity variation (post-1500)	-0.154**	0.205	0.010	0.362
	(0.070)	(0.177)	(0.287)	(0.766)
Land productivity variation (pre-1500)	-0.140**	0.122	-0.044	0.278
	(0.068)	(0.181)	(0.224)	(0.507)
Controls	Yes	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes
Adjusted R^2	0.39	0.49	0.46	0.42
Observations	1497	142	145	108
Sample Restriction	None	$\geq$ 90% Elevation	$\geq$ 90% Ruggedness	New World

Unit of observation: border buffer zone (100km). This table tests the sensitivity of baseline estimates by limiting the full and sibling sample to observations equal or greater than the 90th percentile in elevation (column 2), ruggedness (column 3) and New World observations (column 4). Control variables are identical to the complete set of baseline control variables used in Table 1. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A16: Border-Level Regressions: Mediating Channel – Historical Old World Trade Routes

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0,1)$						
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Full Sample					
Log distance to trade route pre-600 CE	0.004	0.010			-0.000	0.013*
	(0.006)	(0.007)			(0.008)	(0.007)
Log distance to trade route pre-1800 CE			0.007*	0.013***	0.013***	
			(0.004)	(0.004)	(0.005)	
Δ log distance to trade route (pre-1800 $-$ pre-600)						0.013***
						(0.005)
Controls	No	Yes	No	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.29	0.33	0.29	0.33	0.33	0.33
Observations	6205	6205	6205	6205	6205	6205
	Panel B: Sibling Sample					
Log distance to trade route pre-600 CE	0.021*	0.024**			0.011	0.031**
	(0.012)	(0.012)			(0.014)	(0.013)
Log distance to trade route pre-1800 CE			0.020***	0.022***	0.019**	
			(0.008)	(0.008)	(0.009)	
Δ log distance to trade route (pre-1800 $-$ pre-600)						0.019**
						(0.009)
Controls	No	Yes	No	Yes	Yes	Yes
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R^2	0.35	0.37	0.36	0.37	0.37	0.37
Observations	1389	1389	1389	1389	1389	1389

Unit of observation: border buffer zone (100km). This table documents a positive association between the linguistic distance of adjacent ethnic groups and their distance to the nearest pre-600 CE and pre-1800 CE trade route. The trade route data map Old World routes, so both the full sample and sibling sample are restricted to observations located in Old World countries. Control variables are identical to the complete set of baseline control variables used in Table 1. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Computerized Lexicostatistical Similarity

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 (Wichmann et al., 2016). 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.[2]

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).[3] 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 $1 \leq i \leq M$.[4] 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 dissimilarity 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

[2] 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).

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

[4] 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.

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.

C Data Description and Sources

Ethnolinguistic groups: Georeferenced group data comes from the World Language Mapping System (WLMS, 2009). These data map information from each ethnolinguistic group in the *Ethnologue* to the corresponding polygon. When constructing buffer zones are group borders, I use Goode's homolosine map projection.
Source: `http://www.worldgeodatasets.com/language/`

Land productivity: I use the caloric suitability index (CSI) from Galor and Ozak (2016). CSI is a measure of land productivity that reflects the potential caloric output of a grid cell. It's based on the Global Agro-Ecological Zones (GAEZ) project of the Food and Agriculture Organization (FAO). A variety of related measures are available: in the reported estimates I use both the pre-1500 and post-1500 average land productivity measure that includes cells with zero productivity. Land productivity is measured as the average pre-1500 CSI within each spatial unit of observation (border buffer zone or group territory, depending the dataset), and in each case using Goode's homolosine map projection to minimize area distortions. I similarly calculate post-1500 CSI and calculate the change in land productivity as the difference between post-1500 CSI and pre-1500 CSI within each spatial unit of observation. These measures are then converted from millions of kilo calories to thousands of kilo calories by dividing by 1,000.
Source: `http://omerozak.com/csi` and ArcGIS calculations.

Variation in land productivity: I use the caloric suitability index (CSI) from Galor and Ozak (2016). CSI is a measure of land productivity that reflects the potential caloric output of a grid cell. It's based on the Global Agro-Ecological Zones (GAEZ) project of the Food and Agriculture Organization (FAO). Variation in land productivity is measured as the standard deviation of pre-1500 CSI within each spatial unit of observation (border buffer zone or group territory, depending the dataset), and in each case using Goode's homolosine map projection to minimize area distortions. I similarly calculate post-1500 CSI standard deviation, and calculate the change in land productivity in the post-1500 period using Stata. These measures are then converted from millions of kilo calories to thousands of kilo calories by dividing by 1,000.
Source: `http://omerozak.com/csi` and ArcGIS calculations.

Elevation: Elevation data comes from the National Geophysical Data Centre (NGDC) at the National Oceanic and Atmospheric Administration (NOAA, 1999). I use the Goode's homolosine map projection for both the border-level and group-level calculations to minimize area distortion.
Source: `www.ngdc.noaa.gov/mgg/topo/globe.html` and ArcGIS calculations.

Ruggedness: Ruggedness is measured as the standard deviation of the NOAA (1999) elevation data. I use the Goode's homolosine map projection for both the border-level and group-level calculations to minimize area distortion.
Source: https://www.ngdc.noaa.gov/mgg/topo/globe.html and ArcGIS calculations.

Precipitation: Precipitation data comes from the WorldClim (2006) – Global Climate Database, which is based on Hijmans et al. (2005). I measure average precipitation and precipitation variation as the mean and standard deviation, respectively. I use the Goode's homolosine map projection for both the border-level and group-level calculations to minimize area distortion.
Source: http://www.worldclim.org/current and ArcGIS calculations.

Temperature: Temperature data comes from the WorldClim (2006) – Global Climate Database, which is based on Hijmans et al. (2005). I measure average temperature and temperature variation as the mean and standard deviation, respectively. I use the Goode's homolosine map projection for both the border-level and group-level calculations to minimize area distortion.
Source: http://www.worldclim.org/current and ArcGIS calculations.

Malaria Ecology Index: I sourced the Malaria Ecology Index data from Kiszewski et al. (2004). The index measures the prevalence of malaria for each 0.5×0.5 grid cell on earth. I use the Goode's homolosine map projection for both the border-level and group-level calculations to minimize area distortion.
Source: https://sites.google.com/site/gordoncmccord//datasets and ArcGIS calculations.

Log distance between group centroids: I use ArcGIS to calculate the centroid of each adjacent language group pair. I then calculate the great-circle distance between group centroids using the haversine formula. This variable is only relevant to the border-level analysis.
Source: Calculated using ArcGIS and Stata.

Log distance to coast: Georeferenced data on coastlines comes from Natural-Earth (2016). I use the Fuller projection in ArcGIS to minimize distance distortions. For the border-level analysis, I calculate the distance in kilometers from each buffer zone centroid to the nearest coast. For the group-level analysis, I calculate distance from the ethnolinguistic group's centroid. For both analyses, I use the natural log of distance to the coast.
Source: https://www.naturalearthdata.com/downloads/110m-physical-vectors/110m-coastline/
and ArcGIS calculations.

Log distance to border: I use the Digital Chart of the World's georeferenced data on country borders, which comes from the WLMS (2009). I use the Fuller projection in ArcGIS to minimize

distance distortions. For the border-level analysis, I calculate the distance in kilometers from each buffer zone centroid to the nearest national border. For the group-level analysis, I calculate distance from the ethnolinguistic group's centroid. For both analyses, I use the natural log of distance to the border.
Source: http://www.worldgeodatasets.com/language/ and ArcGIS calculations.

Log distance to lake: Georeferenced data on lakes comes from NOAA's (2017) Global Self-consistent Hierarchical High-resolution Geography, Version 2.3.7 June 15, 2017 (Wessel and Smith, 1996). I project the full resolution Level 2 shapefile ("Lakes") into the Fuller projection to minimize distance distortions. For the border-level analysis, I calculate the distance in kilometers from each buffer zone centroid to the nearest lake, and for the group-level analysis, I calculate distance from the ethnolinguistic group's centroid. For both analyses, I use the natural log of distance to the lake.
Source: https://www.ngdc.noaa.gov/mgg/shorelines/ and ArcGIS calculations.

Log distance to major river: Georeferenced data on major rivers comes from NOAA's (2017) Global Self-consistent Hierarchical High-resolution Geography, Version 2.3.7 June 15, 2017 (Wessel and Smith, 1996). I project the full resolution shapefile for river size categories 1-3 into the Fuller projection to minimize distance distortions. For the border-level analysis, I calculate the distance in kilometers from each buffer zone centroid to the nearest major river. For the group-level analysis, I calculate distance from the ethnolinguistic group's centroid. For both analyses, I use the natural log of distance to the major river.
Source: https://www.ngdc.noaa.gov/mgg/shorelines/ and ArcGIS calculations.

Log distance to minor river: Georeferenced data on minor rivers comes from NOAA's (2017) Global Self-consistent Hierarchical High-resolution Geography, Version 2.3.7 June 15, 2017 (Wessel and Smith, 1996). I project the full resolution shapefile for river size categories 4 and 5 into the Fuller projection to minimize distance distortions. For the border-level analysis, I calculate the distance in kilometers from each buffer zone centroid to the nearest minor river. For the group-level analysis, I calculate distance from the ethnolinguistic group's centroid. For both analyses, I use the natural log of distance to the minor river.
Source: https://www.ngdc.noaa.gov/mgg/shorelines/ and ArcGIS calculations.

Total area: For the border-level analysis, I use the natural log of total land area for both ethnolinguistic group homelands, measured in kilometers squared. For the group-level analysis, I use the natural log of total land area for a group's homeland.
Source: Calculated using ArcGIS.

Language population: Ethnolinguistic group population comes from the WLMS *Ethnologue* database (Lewis, 2009). For the border-level analysis, I use the natural log of aggregate population for both groups associated with a buffer zone. In the group-level analysis, I use the natural log of (1 + population) to ensure no observations are dropped. The *Ethnologue* reports contemporary population levels, yet the group-level analysis relies on historical outcome variables, so even though a population is reported as having zero population today, the territory would have been populated by that group in the historical period.
Source: Calculated using Stata.

Latitude and Longitude: For the border-level analysis, latitude and longitude coordinates for an ethnolinguistic group correspond to a group's centroid—not the buffer zone centroid. Differences are calculated by taking the absolute difference of a neighboring pair centroids. For the group-level analysis, I control for latitude and longitude using group centroid coordinates.
Source: Calculated using ArcGIS and Stata.

Historical dependence on agriculture for subsistence: An indicator variable denoting whether a group historically relied on agriculture as their primary means of subsistence or not. Based on a 0-9 scale indexing a group's historical reliance on agriculture (v5). The indicator variable is equal to one if a group is more reliant on agriculture than pastoralism (v4), fishing (v3) or hunting-gathering (v1 + v2). This variable is only relevant to the group-level analysis.
Source: Based on v5 coded in the *Ethnographic Atlas* (Murdock, 1967).

Historical dependence on pastoralism for subsistence: An indicator variable denoting whether a group historically relied on pastoralism as their primary means of subsistence or not. Based on a 0-9 scale indexing a group's historical reliance on pastoralism (v4). The indicator variable is equal to one if a group is more reliant on pastoralism than agriculture (v5), fishing (v3) or hunting-gathering (v1 + v2). This variable is only relevant to the group-level analysis.
Source: Based on v4 coded in the *Ethnographic Atlas* (Murdock, 1967).

Historical dependence on fishing for subsistence: An indicator variable denoting whether a group historically relied on fishing as their primary means of subsistence or not. Based on a 0-9 scale indexing a group's historical reliance on fishing (v3). The indicator variable is equal to one if a group is more reliant on fishing than agriculture (v5), pastoralism (v4) or hunting-gathering (v1 + v2). This variable is only relevant to the group-level analysis.
Source: Based on v3 coded in the *Ethnographic Atlas* (Murdock, 1967).

Historical dependence on hunting-gathering for subsistence: An indicator variable denoting whether a group historically relied on hunting-gathering as their primary means of subsistence

or not. Based on a 0-9 scale indexing a group's historical reliance on hunting (v2) and gathering (v1). After aggregating dependence on hunting and gathering, the indicator variable is equal to one if a group is more reliant on hunting-gathering than agriculture (v5), pastoralism (v4) or fishing (v3). This variable is only relevant to the group-level analysis.
Source: Based on v1 and v2 coded in the *Ethnographic Atlas* (Murdock, 1967).

Historical reliance on trade as a means of subsistence: An indicator variable denoting whether a group was historically reliant on any amount of trade as a means of subsistence or not. For all non-missing observations, the indicator takes a value of one if encoded as a "Co-dominant with one or more other categories" or "Important, but not a major subsistence activity" or "Present, but relatively unimportant," and takes a value of zero otherwise. This variable is only relevant to the group-level analysis.
Source: Based on v732 coded in the *Standard Cross-Cultural Sample* (Murdock and White, 1969).

Historical reliance on trade as a food source: An indicator variable denoting whether a group was historically reliant on trade as a source of food or not. For all non-missing observations, the indicator takes a value of one if encoded as an "$>$ 50 pct of food" or "$<$ 50 pct of food, and less than any single local source" or "$<$ 10 pct of food (90 pct form local extractive sources)," and takes a value of zero otherwise. This variable is only relevant to the group-level analysis.
Source: Based on v6 coded in the *Standard Cross-Cultural Sample* (Murdock and White, 1969).

Exogamy (*Ethnographic Atlas*): An indicator variable denoting whether a group's historical community marriage organization practices can be characterized as exogamous or not. For all non-missing observations, the indicator takes a value of one if encoded as an "exogamous community." This variable is only relevant to the group-level analysis.
Source: Based on v15 coded in the *Ethnographic Atlas* (Murdock, 1967).

Exogamy (*SCCS*): An indicator variable denoting whether a group's historical community marriage organization practices can be characterized as exogamous or not. For all non-missing observations, the indicator takes a value of one if encoded as an "Local endogamy 0-10 pct (exogamy)." This variable is only relevant to the group-level analysis.
Source: Based on v72 coded in the *Standard Cross-Cultural Sample* (Murdock and White, 1969).

Conflict: frequently attacks others: An indicator variable denoting whether a group historically attacked other groups or not. For all non-missing observations, the indicator takes a value of one if encoded as an "Continual" or "Frequent" engagement in external wars as the attacker. This variable is only relevant to the group-level analysis.
Source: Based on v892 coded in the *Standard Cross-Cultural Sample* (Murdock and White, 1969).

Conflict: frequently attacked by others: An indicator variable denoting whether a group historically was attacked by other groups or not. For all non-missing observations, the indicator takes a value of one if encoded as an "Continual" or "Frequent" engagement in external wars as the target of attack. This variable is only relevant to the group-level analysis.
Source: Based on v893 coded in the *Standard Cross-Cultural Sample* (Murdock and White, 1969).

Distance to nearest Old World trade route (pre-600 and pre-1800 CE): Digitized maps of Old World trade routes come from Michalopoulos et al. (2016, 2018). Great-circle distance is calculated between border buffer zone centroid and the nearest Old World route using the haversine formula. This variable is only relevant to the border-level analysis.
Source: Calculated using ArcGIS and Stata.

References

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy*, 113(1):151–184.

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.

Bondarenko, D., Kazanov, A., Khaltourina, D., and Korotayev, A. (2005). Ethnographic Atlas XXI: Peoples of Easternmost Europe. *Ethnology*, 44:261–289.

Colella, F., Lalive, R., Sakalli, S. O., and Thoenig, M. (2019). Inference with Arbitrary Clustering. *IZA Discussion Paper 12584*, (Vcv):1–15.

Galor, O. and Ozak, O. (2016). The Agricultural Origins of Time Preference. *American Economic Review*, 106(10):3064–3103.

Giuliano, P. and Nunn, N. (2018). Ancestral Characteristics of Modern Populations. *Economic History of Developing Regions*, 33(1):1–17.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very High Resolution Interpolated Climate Surfaces for Global Land Areas. *International Journal of Climatology*, 25(15):1965–1978.

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.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., and Sachs, J. (2004). A Global Index Representing the Stability of Malaria Transmission. *American Journal of Tropical Medicine and Hygiene*, 70(5):486–498.

Korotayev, A., Kazankov, A., Borinskaya, S., Khaltourina, D., and Bodarenko, D. (2004). Ethnographic Atlas XXX: Peoples of Siberia. *Ethnology*, 43:83–92.

Lewis, M. P. (2009). *Ethnologue: Languages of the World (Sixteenth Edition)*. SIL International, Dallas, Texas.

Michalopoulos, S., Naghavi, A., and Prarolo, G. (2016). Islam, Inequality and Pre-Industrial Comparative Development. *Journal of Development Economics*, 120:86–98.

Michalopoulos, S., Naghavi, A., and Prarolo, G. (2018). Trade and Geography in the Spread of Islam. *Economic Journal*, 128(616):3210–3241.

Murdock, G. P. (1967). *Ethnographic Atlas*. University of Pittsburgh Press, Pittsburgh.

Murdock, G. P. and White, D. R. (1969). Standard Cross-Cultural Sample. *Ethnology*, 9:329–369.

Natural-Earth (2016). Natural Earth 1:10m Physical Vectors Coastline.

NOAA (1999). *The Global Land One-Kilometer Base Elevation (GLOBE) Digital Elevation Model (Version 1)*. National Oceanic and Atmospheric Administration, National Geophysical Data Center, Boulder.

NOAA (2017). *Global Self-Consistent, Hierarchical, High-Resolution Geography Database (GSHHG) Version 2.3.7*. National Oceanic and Atmospheric Administration, National Geophysical Data Center, Boulder.

Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business Economics and Statistics*, 37(2):187–204.

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.

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.

Wessel, P. and Smith, W. H. F. (1996). A Global Self-consistent, Hierarchical, High-resolution Shoreline Database. *Journal of Geophysical Research*, 101(B4):8741–8743.

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

Wichmann, S., Holman, E. W., and Brown, C. H. (2016). *The ASJP Database (Version 17)*.

WLMS (2009). *World Language Mapping System (Version 16)*. SIL International, Dallas, Texas.

WorldClim (2006). *WorldClim 1.4 Climate Data for 1960-1990*.