Essays on the Economics of Ethnolinguistic Differences

Andrew C. Dickens

A DISSERTATION

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN ECONOMICS
YORK UNIVERSITY
TORONTO, ONTARIO
APRIL 2017

Abstract

In this dissertation, I study the origins and economic consequences of ethnolinguistic differences. To quantify these differences, I construct a lexicostatistical measure of linguistic distance. I use this measure to study two different outcomes: ethnic politics and cross-country idea flows. I then take the economic importance of ethnolinguistic differences as given, and explore the geographic foundation of these differences.

In chapter 1, I document evidence of ethnic favoritism in 35 sub-Saharan countries. I use lexicostatistical distance to quantify the similarity between an ethnic group and the national leader's ethnic identity. I find that a one standard deviation increase in similarity yields a 2 percent increase in group-level GDP per capita. I then use the continuity of lexicostatistical similarity to show that favoritism exists among groups that are not coethnic to the leader, where the mean effect of non-coethnic similarity is one quarter the size of the coethnic effect. I relate these results to the literature on coalition building, and provide evidence that ethnicity is a guiding principle behind high-level government appointments.

In chapter 2, I use book translations data to capture cross-country idea flows. It has been conjectured that income gaps are smaller between ancestrally related countries because they communicate more ideas. I provide empirical support for this link and a deeper understanding of the hypothesized mechanism: population differences do exhibit a negative relationship with the diffusion of ideas, with the caveat that this negative relationship operates across linguistic lines. After accounting for the linguistic distance between two countries, I find that dissimilar populations communicate more ideas.

In chapter 3, I study the geographic origins of ethnolinguistic differences. I construct a novel dataset to examine the border regions of neighbouring ethnolinguistic groups, together with variation in the set of potentially cultivatable crops at the onset of the Columbian Exchange, to estimate how agricultural diversity impacts linguistic differences between neighbouring groups. I find that ethnic groups separated across agriculturally diverse regions are more similar in language than groups separated across homogeneous agricultural regions. I propose that historical trade in agriculturally diverse regions is the mechanism by which group similarities are preserved.

Dedication

To Meghan, for your steadfast encouragement and support throughout the years. You have made my life better in more ways than I could ever articulate to you, and undoubtedly my work is better off as a result. Your daily inspiration is the footing upon which I have found my stride as a researcher.

To my parents, Wenda and Greg, you have instilled in me an appreciation of knowledge. Thank you for believing in me, and always encouraging me to go after my dreams in life – academic and otherwise. My education would not have been possible without the opportunities that you provided for me, and I am forever grateful that I am where I am today because of you.

To my sister, Claire, I have always admired your curiosity and enthusiasm. You have inspired these qualities in me, without which I would not have seen this dissertation through to completion. You have always been my ally and I thank you for your unwavering support.

Acknowledgements

I am indebted to my supervisor, Nippe Lagerlöf, for his unwavering support of my ideas, for his ability to always push me towards my intellectual frontier, and for the countless hours of his time he spent with me discussing my research. My dissertation would pale in comparison without his invaluable supervision, his contribution to my academic development and his friendship.

My development as a researcher would also not have been the same without the help of Tasso Adamopoulos. Throughout many meetings over the course of my dissertation, he has offered perspective and encouragement that has directly contributed to how I think about economic problems. I am truly grateful for all the excellent advice Tasso has given me about my research and about how to succeed as an economist.

I owe a huge thanks to Ben Sand for guiding me through the world of econometrics. My understanding and appreciation of empirical methods would not be the same without Ben's wealth of knowledge. In addition to his fantastic comments and suggestions about my work, Ben has spent hours of his time helping me understand the answer to my many questions over the years.

Many other members of the department have contributed to my dissertation and provided an environment that was conducive to my academic development. A special thanks to Berta Esteve-Volart for reading and commenting on many drafts of my work, and for always providing excellent comments during or after one of my seminars. Other faculty members that have contributed to my dissertation and always given me their time, even though not on my committee, include Ahmet Akyol, Sam Bucovetsky, Avi Cohen, Wai-Ming Ho, Fernando Leibovici, Uros Petronijevic, Laura Salisbury and Andrey Stoyanov. And a big thanks to my classmate, Andrew Hencic, for always hearing out my daily research problem on the subway ride home from campus.

An additional thanks to the many discussants and seminar participants that have contributed to my dissertation, in particular James Fenske for providing extensive comments on my job market paper. A special thanks also goes to Oded Galor for hosting my Visiting Research Fellowship at Brown University in 2015. My time at Brown was the most memorable experience of my graduate studies, and I thank those that I met who contributed to my dissertation, including Greg Casey, Mario Carillo, Raphaël Franck, Stelios Michalopoulos, Ömer Özak, Assaf Sarid and David Weil.

Last, but definitely not least, I am thankful for my friends who have made Toronto home for me. Having a group of friends who support and encourage me – away from work – has been instrumental to my productivity and focus over the past six years. This dissertation would not be the same without them.

Table of Contents

List of Tables

List of Figures

Chapter 1

Ethnolinguistic Favoritism in African Politics

1.1 Introduction

Understanding why global poverty is so concentrated in Africa remains one of the most crucial areas of inquiry in the social sciences. One long-standing explanation is that Africa's high level of ethnic diversity is a major source of its underdevelopment and political instability (Easterly and Levine, 1997; Collier and Gunning, 1999; Posner, 2004; Alesina and La Ferrara, 2005, among others). Yet recent evidence documents that the source of underdevelopment is not ethnic diversity per se, but rather Africa's high degree of inequality between ethnic groups (Alesina et al., 2016). This suggests that ethnic diversity is only an impediment to economic development when some ethnicities prosper at the expense of others.

Ethnic inequality not only contributes to the under-provision of the overall level of public resources (Baldwin and Huber, 2010), but it provokes discriminatory policies that advantage some groups over others (Alesina et al., 2016). Discriminatory policies of this type are a form of ethnic favoritism, which has been the subject of a few influential papers that document evidence of public resource distribution across ethnic lines in Africa (Franck and Rainer, 2012; Burgess et al., 2015; Kramon and Posner, 2016). The provision of resources on the basis of ethnicity – rather than on a need or marginal value basis – suggests that some ethnic groups are being systematically favored over others. Hence, a better understanding of how ethnic patronage is distributed and to whom is important because it sheds light on the extent to which favoritism occurs and how some ethnic groups benefit at the expense of others.

In this paper I revisit the study of ethnic favoritism with three contributions. My first contribution is a novel measure of linguistic similarity that quantifies the relative similarity of all ethnic groups to the national leader in each country, not just groups that share an ethnicity with the leader (coethnics). Because linguistic similarity is measured on the unit interval it encompasses the commonly used coethnic dummy variable, while extending measurement to all non-coethnic

groups.[1] This extension is beneficial because the majority of Africans are never coethnic to their leader.[2] The continuity of this new measure implies that *any* change in the ethnic identity of a leader is associated with *some* change in the similarity of *all* groups in a country, an important source of variation that is not observable using a coethnic dummy variable. This measure also provides testable grounds for the central hypothesis of this paper: a group's well-being is increasing in their ethnic similarity to the national leader.

My second contribution relates to the evidence that ethnic favoritism is widespread throughout sub-Saharan Africa. I use the systematic partitioning of African ethnic groups across political borders to expand the scope of evidence relative to previous studies. In particular, I exploit the fact that the same ethnic group is split between neighboring countries and exposed to a different ethnic leader on each side of the border. As different ethnic leaders come and go from power, the relative similarity of a partitioned group varies over time. This source of variation allows for ethnicity-year fixed effects, a novel empirical specification that accounts for the long-run persistence of a group's pre-colonial history on group-level outcomes today (Gennaioli and Rainer, 2007; Michalopoulos and Papaioannou, 2013; Fenske, 2013). I use this variation in a triple difference set-up and document evidence of ethnic favoritism in two empirical settings: a panel of 163 ethnic groups split across 35 countries, and a repeated cross-section of individuals living in 20 groups split across 13 countries. I also use the continuity of similarity in both settings to show that favoritism exists among groups that are not coethnic to the leader, a new finding that is a contribution in itself. This speaks to why a continuous measure of ethnic similarity is important: ethnic favoritism is under-reported when using a coethnic dummy variable because non-coethnic favoritism goes undetected. In order to understand the impact of ethnic favoritism on development, it is important to understand the extent to which it occurs.

For my third contribution I disentangle the relative importance of location-based favoritism from individual-level favoritism. It is commonly assumed that the ethnic majority of a region defines the ethnic identity of that region, despite the fact that not all residents belong to the dominant group. While this is a reasonable assumption, it limits our understanding of how patronage is distributed because no distinction can be made between regional transfers and targeted transfers towards individuals. To understand who benefits from favoritism it is necessary to understand whether the benefits of similarity are exclusive to individuals living in their ethnic homeland, or if patronage is distributed more broadly by targeting individuals irrespective of location. To this end, I use variation among survey respondents who identify with an ethnicity that is different from the ethnic region in which they live.[3] I find that patronage is distributed according to the ethnic identity of a region rather than as a targeted transfer towards individuals from a particular

[1] People identify as coethnics because they share a common ancestry and language, hold similar cultural beliefs and pursue related economic activities (Batibo, 2005). In this way, linguistic similarity is a good measure of ethnic proximity because it is the most visible marker of ethnic identity.

[2] Using population data from the Ethnologue (16^{th} edition), I calculate that only 34 percent of the median sub-Saharan country's population was *ever* coethnic to their leader between 1992 and 2013.

[3] This is analogous to Nunn and Wantchekon (2011), who use a similar source of variation to separate internal norms of an individual from the external norms of an individual's environment.

ethnic group.

Throughout this empirical analysis I rely on the fact that the location of African borders are quasi-random (Englebert et al., 2002; Michalopoulos and Papaioannou, 2014, 2016, among others). The historical formation of Africa's borders began with the Berlin Conference of 1884-1885, where European powers divided up Africa with little regard for the spatial distribution of ethnic homelands (Herbst, 2000). This disregard led to the arbitrary formation of national borders, which "did not reflect reality but helped create it" (Wesseling, 1996, p.364). One such reality was the partitioning of approximately 200 ethnic groups throughout Africa.

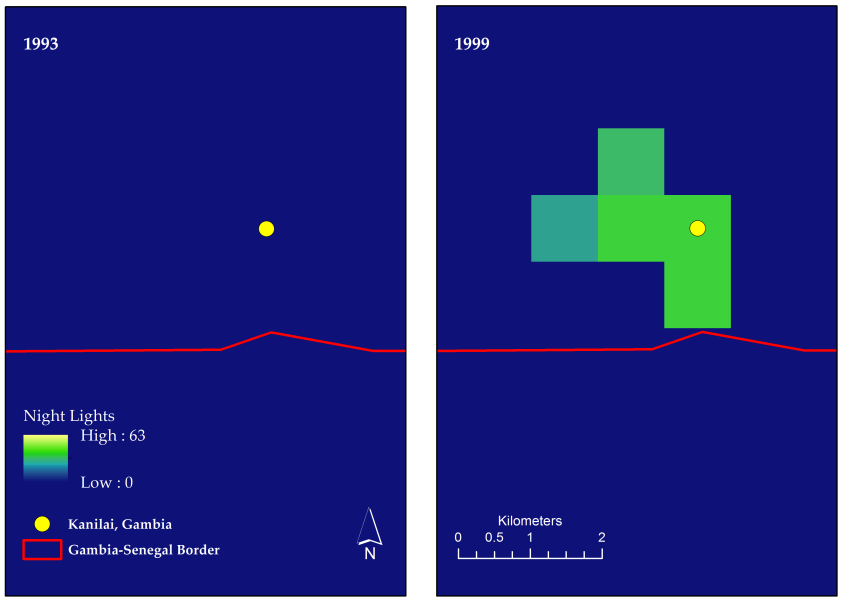
In the context of this study, the quasi-random nature of African border design generates exogenous variation because the ethnic identity of a national leader varies across borders within the same partitioned group. Because an ethnic group shares a common ancestry, and is relatively homogeneous in terms of cultural and biological factors, the fraction of a partitioned group on one side of the border is a suitable counterfactual observation for the other fraction of that same group across the border. Using ethnic groups partitioned across African borders as a source of exogenous variation is methodologically similar to Michalopoulos and Papaioannou (2014).

To exploit this within-group variation I use the 16^{th} edition of the Ethnologue language map (Lewis, 2009). This map depicts the spatial distribution of ethnolinguistic homelands across the world. I use these subnational groups as a spatial unit of observation in Africa. Because no income data exists at this level of observation, I proxy an ethnic group's economic activity using annual satellite images of night light luminosity for the time period 1992-2013. These luminosity data are available at a very fine spatial resolution, which I can use to construct a panel of economic activity at the country-group level. Luminosity is frequently used as a measure of economic activity because of its strong empirical association with GDP per capita and other measures of living standards (Henderson et al., 2012; Michalopoulos and Papaioannou, 2013, 2014; Alesina et al., 2016, among others). Hodler and Raschky (2014) first used luminosity in this way to study patterns of regional favoritism.

Consider, as an example, the Jola-Fonyi language group partitioned across Gambia and Senegal. In 1993, both the Gambian and Senegalese Jola-Fonyi bear little resemblance to their respective leaders. For several years little changed in Senegal as President Diouf's reign continued. On the contrary, much changed for the Gambian Jola-Fonyi when Yahya Jammeh, a young officer in the National Gambian Army, overthrew President Jawara in a 1994 military coup. Jammeh was born in Kanilai, a small village near the southern border of Gambia and home to the Jola-Fonyi language group. Jammeh took much pride in his birth region – a "place that gained prominence overnight in Gambia" (Mwakikagile, 2010, p. 56). Jammeh repeatedly "feathered his nest" to such an extent that the Jola-Fonyi region surrounding Kanilai is one of few rural areas in Gambia with "electricity, street lighting, paved roads and running water – not to mention its own zoo and game preserve, wrestling arena, bakery and luxury hotel with a swimming pool" (Wright, 2015, p. 219).

Figure 1.1 provides visual evidence of this phenomenon. The two panels represent the same subsection of the Jola-Fonyi language group at two points in time, with the border dividing Gam-

Figure 1.1: Change in Night Lights Intensity from 1993-1999



This figure documents the change in night light activity in the partitioned Jola-Fonyi language group in Gambia (north of the border) and Senegal (south of the border) between 1993 and 1999. In 1994, Yahya Jammeh assumed power of Gambia and soon after started reallocating funds to the Jola-Fonyi. Within 5 years of presidency the Gambian Jola-Fonyi exhibit much greater economic activity in terms of night lights than the Senegalese Jola-Fonyi on the south side of the border, whom had no change in leadership during this period.

bia to the north and Senegal to the south. While there is no visible night light activity on either side of the border in 1993, there is a significant increase in lights on the Gambian side only 5 years after Jammeh assumed power. On the contrary, Diouf's presidency continued throughout this entire period and there is no observable change in night light activity in Senegal just south of the border. This demonstrated change in Figure 1.1 is exactly the within-group variation that I use to estimate the effect of similarity. In this case, the Senegalese Jola-Fonyi are the counterfactual observation for the Gambian Jola-Fonyi, who are equally dissimilar in language to their incumbent leader in 1993, and the effect of similarity on night light activity is estimated off of the change in linguistic similarity following Jammeh's rise to power.

My benchmark results imply that a standard deviation increase in linguistic similarity (23 percent) yields a 7 percent increase in luminosity and a 2 percent increase in group-level GDP per capita. I also use the continuity of linguistic similarity to document evidence of non-coethnic favoritism, where the mean non-coethnic effect is one quarter the coethnic premium. To the contrary I find no evidence of anticipatory effects in the data or evidence migration in response to leadership changes. To be sure this result is not a consequence of my new measure of similarity I construct two alternative measures: a standard binary measure of coethnicity and a discrete similarity measure of the ratio of shared nodes on the Ethnologue language tree. While these alternative measures of similarity yield significant evidence of favoritism, my preferred lexicostatistical measure of similarity is more precisely estimated and the only measure to maintain significance

in a series of horse race regressions.

I also test for a variety of mechanisms, but find no systematic evidence of the usual channels (e.g., democracy). However, I do find that my benchmark result is largely driven by leaders who have held office longer than the sample median of nine years. This implies that one determinant of favoritism is leadership tenure.

Next I turn to individual-level data from the Demographic and Health Survey (DHS). I use survey cluster coordinates to pinpoint the location of individual respondents on the Ethnologue map. Doing so allows me to construct a repeated cross-section of individuals living in partitioned ethnic groups across DHS survey waves. Narrowing the focus to these individuals allows me to exploit the same variation I use in my benchmark estimates. As an outcome I use an individual-level measure of access to public resources and ownership of assets. I corroborate my benchmark findings with this individual-level data, including evidence of non-coethnic favoritism. I also establish that patronage is distributed regionally and not as a targeted transfer towards individuals.

These findings speak to a sparse but growing body of evidence that ethnic favoritism is widespread throughout sub-Saharan Africa. Franck and Rainer (2012) use a panel of ethnic groups in 18 countries to document evidence of favoritism throughout sub-Saharan Africa. What sets my paper apart from Franck and Rainer's (2012) is that I construct a panel of *partitioned* ethnic groups, so I have a minimum of two country-group observations for any partitioned group in a year. This feature of my data affords me ethnicity-year fixed effects. Because I can account for all observable and unobservable time-varying features of an ethnic group, I am able to rule out endogeneity concerns associated with the impact of pre-colonial group characteristics on contemporary development (Gennaioli and Rainer, 2007; Michalopoulos and Papaioannou, 2013; Fenske, 2013).

More commonly researchers focus on a single patronage good in a single country. Kramon and Posner (2016) find that Kenyans whom are coethnic to their leader attain higher levels of education, while Burgess et al. (2015) find that Kenyan districts associated with the leader's ethnicity receive two times the investment in roads during periods of autocracy. At an even finer level, Marx et al. (2015) document evidence of ethnic favoritism in housing markets within a large slum outside of Nairobi.

The rich micro-data these studies use provide clear evidence of ethnic favoritism and the channels through which patronage is distributed. Yet generalizing these results is difficult because of the highly localized analyses these studies employ. To this end, I exploit the systematic partitioning of ethnic groups in Africa to expand the scope of evidence to 35 sub-Saharan countries. In a related manuscript, De Luca et al. (2015) document that ethnic favoritism is an axiom of politics on a global scale and not simply an African phenomenon. As this literature continues to grow, these localized studies coupled with the broader evidence of ethnic favoritism help to build consensus around ethnic favoritism in Africa.[4]

[4]Yet consensus on ethnic favoritism if Africa has not been reached. Francois et al. (2015) document that leaders only provide a small premium to their coethnics, and otherwise political power is proportional to group size in Africa. Kasara (2007) finds that leaders are more likely to extract taxes from their own ethnic group because they have a better understanding of internal markets in their homeland.

The notion that ethnic favoritism drives discriminatory policies that disadvantage some groups at the expense of others also relates this research to the literature on ethnic inequality and conflict. Alesina et al. (2016) document that the negative correlation between ethnic inequality and economic development is a global phenomenon, though most pronounced in Africa. Income differences between a country's ethnic groups can also impact the political process: ethnic inequality mitigates public good provision (Baldwin and Huber, 2010), diminishes the quality of governance (Kyriacou, 2013), and provokes the "ethnification" of political parties (Huber and Suryanarayan, 2014). At the heart of this literature is the long-standing instrumentalist view that conflict over scarce resources drives ethnic competition in Africa (Bates, 1974). Even the perception of ethnic favoritism exacerbates already existing ethnic tensions (Bowles and Gintis, 2004), which itself can further incite ethnic conflict (Esteban and Ray, 2011; Esteban et al., 2012; Caselli and Coleman, 2013).

I contribute to this line of research with evidence that regions, rather than individuals, tend to be targeted, and that non-coethnic groups that are similar to the leader stand to gain from their ethnic proximity. The fact that similar but not identical ethnic regions benefit from patronage suggests that ethnicity is more than just a marker of identity: similarity may create affinity or reduce coordination costs across related non-coethnic groups. This is consistent with the idea that these broader ethnic connections may solve collective action problems (Miguel and Gugerty, 2005) and bring about greater between-group trust (Habyarimana et al., 2009). The continuity of linguistic similarity captures these affinities that are otherwise unobservable with a coethnic dummy variable, thus highlighting one further benefit of this new measure.

This is also in line with the idea that leaders bring elites from outside of their ethnic group into the governing coalition in an effort to sustain power in the face of political instability (Joseph, 1987; Francois et al., 2015). In the discussion section of this paper I provide evidence that leaders appoint similar but not identical ethnic elites to high-level government positions for this purpose. In doing so, non-coethnic groups gain coethnic representation in government, where representatives speak on their behalf and channel resources to them (Arriola, 2009). Although my focus is Africa, this deeper understanding of where favoritism is expected to take place has implications for distributive politics more broadly: it contributes to our knowledge of how targeted transfers can potentially magnify inequality between groups and thus is informative of a determining factor of comparative economic development

The rest of this paper is structured as follows. Section 1.2 describes how I identify language group partitions and measure linguistic similarity. This section also documents patterns in the data. Section 1.3 outlines the empirical model and identification strategy, and Section 1.4 reports the benchmark estimates and robustness checks. Section 1.5 disentangles the relative importance of location-based favoritism from individual-level favoritism, and in Section 1.6 I link the findings to the literature on ethnic favoritism and coalition building, and provide suggestive evidence that non-coethnic favoritism works through the appointment of elites from outside of the leader's ethnic group. Section 1.7 concludes.

Figure 1.2: Language Groups

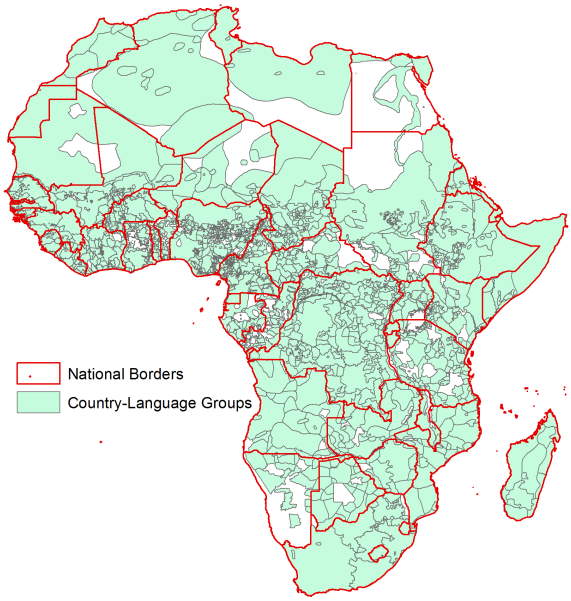
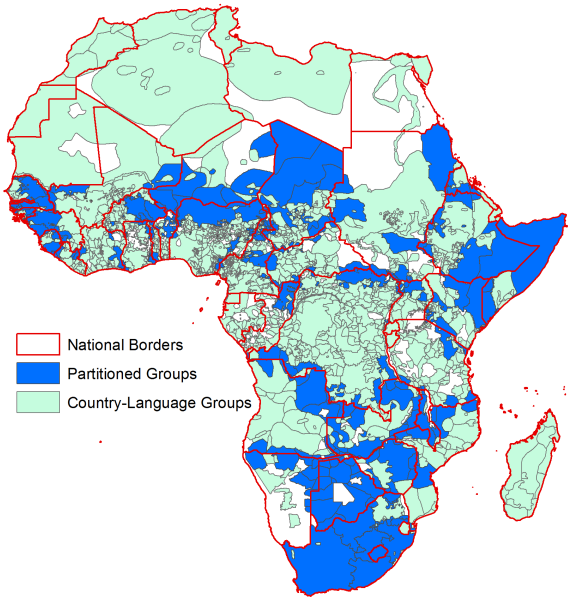


Figure 1.3: Language Partitions



1.2 Data

In this section I describe the main variables of interest. For a complete description of all data and sources see Appendix B.

1.2.1 Language Group Partitions

I construct language group partitions using the 2009 Ethnologue (16^{th} edition) mapping of language groups from the World Language Mapping System (WLMS). These WLMS data depict the spatial distribution of linguistic homelands at the country-language group level (Figure 1.2). I focus on continental sub-Saharan Africa.[5] In total there are 2,384 country-language group observations reflecting 1,961 unique language groups in 42 continental African countries.[6]

I define a partition as a set of contiguous country-language group polygons, where each polygon in a set is part of the same language group but separated by a national border. I use ArcGIS to identify these partitioned groups, excluding country-language groups with a reported Ethnologue population of zero. The result is 486 remaining country-language group observations, made up of 227 language groups partitioned across 37 African countries.

[5] I use the United Nations classification of sub-Saharan countries. However, I include Sudan in the analysis because it is geographically part of sub-Saharan Africa and contains a number groups partitioned between Sudan and sub-Saharan countries.

[6] Because Western Sahara is a disputed territory I exclude it from this border analysis.

1.2.2 Satellite Imagery of Night Light Luminosity

Satellite imagery of night light luminosity come from the National Oceanic and Atmospheric Administration's (NOAA) National Geophysical Data Center. Many others have used these data because of two features: night lights data exhibit a strong empirical relationship with GDP per capita and other measures of living standards (Henderson et al., 2012), and because these data are available at a spatial resolution of 30-arc seconds (approximately 1 square kilometre).[7] The fine resolution of these lights data facilitates a proxy measure of GDP per capita at any desired level of spatial aggregation. Because I require a measure of economic activity at the country-language group level – a level of aggregation where no official data on economic output exists – the availability of these data is indispensable to this study.

The yearly composite of night light luminosity is constructed by NOAA using daily images taken from U.S. Department of Defense weather satellites that circle the earth 14 times a day. These satellites observe every location on earth every night sometime between 20:30 and 22:00. Before distributing these data publicly, NOAA scientists remove observations contaminated by strong sources of natural light, e.g., the summer months when the sun sets late, light activity related to the northern and southern lights, forest fires, etc. All daily images that pass this screening process are then averaged for the entire year producing a satellite-year dataset for the time period 1992 to 2013. Light intensity receives a value of 0 to 63 at a resolution of 30-arc seconds. The result is a measure of night light intensity that only reflects human (economic) activity.[8]

Using these data I construct a panel of average luminosity for each country-language group partition. I use the Africa Albers Equal Area Conic projection to minimize distortion across the area dimension before calculating the average light luminosity of each country-language group polygon in each year.[9] I follow Michalopoulos and Papaioannou (2013, 2014) and Hodler and Raschky (2014) in adding 0.01 to the log transformation of the lights data because roughly 40% of these data have a value of zero in the benchmark sample. Doing so helps correct for the non-normal nature of the data and preserves sample size, and allows for a (near) semi-elasticity interpretation of the benchmark empirical model.

1.2.3 Assignment of a Leader's Ethnolinguistic Identity

There are 106 leaders to assign an ethnolinguistic identity for my sample of 35 countries between 1992-2013. The challenge of mapping ethnicity to language is that, in some instances, a single ethnic group speaks many languages. Because African language groups are often resident of well-defined territories (Lewis, 2009), an ethnolinguistic identity is typically attached to a person's

[7] Hodler and Raschky (2014) also show there is a strong empirical relationship between these night lights data and GDP at the subnational administrative region. Michalopoulos and Papaioannou (2014) further validate the use of night lights in Africa as a proxy measure of development with evidence that light intensity correlates strongly with individual-level data on electrification, presence of sewage systems, access to piped water and education.

[8] Henderson et al. (2012, p. 998) provide a thorough introduction to the NOAA night lights data.

[9] In some years data is available for two separate satellite, and in all such cases the correlation between the two is greater than 99 percent in my sample. To remove choice on the matter I use an average of both.

birthplace (Batibo, 2005). As a first step towards assignment I locate the birthplace of a leader and collect latitude and longitude coordinates for each birthplace from www.latlong.net. I project these coordinates onto the Ethnologue map of Africa to back out the language group associated with each leader's birthplace.[10] I exclude leaders born abroad (4 leaders) since their ethnolinguistic group is not home to the country they govern.[11] Second, I identify a leader's ethnic identity using data from Dreher et al. (2015) and Francois et al. (2015), and in the few instances where neither source reports the ethnicity of a leader I fill in the gap using a country's Historical Dictionary. Finally, I take the following steps to assign a leader's ethnolinguistic identity using these data:

Step 1: I compare the birthplace linguistic identity with the ethnic identity for each of the 102 leaders. In 56.9 percent of the sample the name of the birth language and ethnic identity are equivalent (58 leaders). For these leaders the assignment is unambiguous.

Step 2: For the remaining sample of unmatched leaders, I check if the birthplace language is a language spoken by the leader's ethnic group. In 12.7 percent of the sample this is true (13 leaders); I assign the birthplace language as the leader's ethnolinguistic identity.

Step 3: For the remaining 30.4 percent of unmatched leaders (31 leaders), the birthplace identity does not correspond to their ethnic identity. This is especially true for leaders born in a major city. For these leaders I drop the birthplace identity and map the ethnicity of a leader to a single language using the three-step assignment rule outlined in Appendix B.

1.2.4 Linguistic Similarity

Estimating linguistic similarity is difficult because languages can differ in a variety of ways, including vocabulary, pronunciation, grammar, syntax, phonetics and more. One common approach is to use a measure of the shared branches on a language tree as an approximation of linguistic similarity. Known as cladistic similarity, this measure was introduced to economists by Fearon and Laitin (1999), popularized by Fearon (2003) and has since become the convention.[12] The idea behind the cladistic approach is that two languages with a large number of shared nodes – and thus a recent splitting from a common ancestor – will be similar in terms of language because of their common ancestry. The data most commonly used is Fearon's (2003) cladistic measure of linguistic similarity, constructed using the Ethnologue's phylogenetic language tree. A cladistic measure is attractive because linguistic similarity is easily computed for any language pair, since language trees exist for virtually all known world language families (Lewis, 2009). See Appendix A for a formal definition of this measure.

My preferred measure is a computerized lexicostatistical measure of linguistic similarity de-

[10] Because most leaders enter/exit office mid-year, I assign the incumbent leader as whomever is in power on December 31st of the transition year. Hence, by assumption I drop any leader who exited office the same year she entered office because she was neither in power the previous year or December 31st of the transition year.

[11] These leaders include Ian Khama (Botswana), Francois Bozize Yangouvonda (Central African Republic), Nicephore Soglo (Benin), and Rupiah Banda (Zambia).

[12] For example, Guiso et al. (2009); Spolaore and Wacziarg (2009); Desmet et al. (2012); Esteban et al. (2012) and Gomes (2014) all use a cladistic approach, among others.

veloped by the Automatic Similarity Judgement Program (ASJP).[13] As a percentage estimate of a language pair's cognate words (i.e., words that share a common linguistic origin), the lexicostatistical method is a measure of the phonological similarity between two languages. Hence, a lexicostatistical measure can be thought of as a proxy for the ancestral relationship between two groups, or an implicit measure of the set of shared ancestral and cultural traits that are important to group identity.

The ASJP Database (Version 16) consists of 4401 language lists, where each list contains the same 40 implied meanings (i.e., words) for comparison across languages. The ASJP research team has transcribed these lists 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. Meanings are then transcribed according to pronunciation before language differences are estimated.[14]

Then for each language pair of interest I run the Levenshtein distance algorithm on the respective language lists, which calculates the minimum number of edits necessary to translate the spelling of each word from one language to another. To correct for the fact that longer words will demand more edits, each distance is divided by the length of the translated word. This normalization yields a percentage estimate of dissimilarity, which is measured across the unit interval. The average distance of a language pair is calculated by averaging across the distance estimates of all 40 words. By this procedure I estimate the linguistic distance of a language pair vis-à-vis the vocabulary dimension.

A second normalization procedure is used to adjust for the accidental similarity of two languages (Wichmann et al., 2010). This normalization accounts for similar ordering and frequency of characters that are the result of chance and independent of a word's meaning. Finally, I define the lexicostatistical similarity of a language pair as one minus this normalized distance. For a formal definition of this measure, I direct to reader to Appendix A.

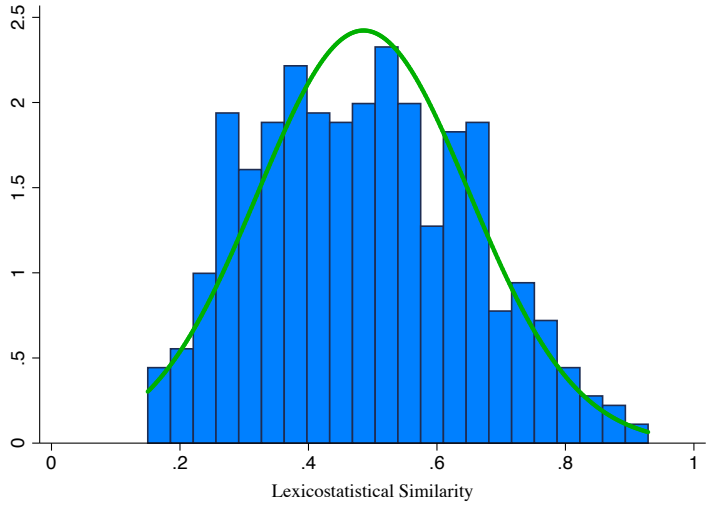
The main advantage of the lexicostatistical approach is that it measures similarity in a more continuous way than the cladistic approach. Because the lexicostatistical method explicitly identifies the phonological differences of a language pair, there is far more observable variation in a measure of lexicostatistical similarity than cladistic similarity. The cladistic approach is a coarse measure of similarity because data dispersion is limited to 15 unique values, the maximum number of language family classifications in the Ethnologue.

To illustrate this point, consider language pairs that share a common parent language on the Ethnologue language tree. Let these language pairs be known as siblings. All sibling pairs share the maximum number of tree nodes, and have no differences in cladistic similarity between them, but they do exhibit substantial variation in lexicostatistical similarity. To make this point clear,

[13]The lexicostatistical measure I use in this paper has been used to study factor flows in international trade (Isphording and Otten, 2013), job satisfaction of linguistically distinct migrants (Bloemen, 2013), language acquisition of migrants (Isphording and Otten, 2014), and the role of language in the flow of ideas (Dickens, 2016b). See (Ginsburgh and Weber, 2016) for a discussion of this and other measures of linguistic distance.

[14]For example, the French word for *you* is *vous*, and is encoded using ASJPcode as *vu* to reflect its pronunciation.

Figure 1.4: Lexicostatistical Similarities Among Sibling Language Pairs



This figure establishes the additional variation introduced by a lexicostatistical measure of linguistic similarity that is not observable with a cladistic measure of similarity. The histogram plots the estimates of lexicostatistical similarity among sibling language pairs for all of Africa ($n = 1,241$). Sibling language pairs are those that share a parent language on the Ethnologue language tree, which by definition implies that among sibling language pairs there is no observable variation in cladistic similarity.

I plot the distribution of lexicostatistical similarities among all African sibling language pairs in Figure 1.4. This highlights the sizeable dispersion in lexicostatistical similarities among sibling language pairs.

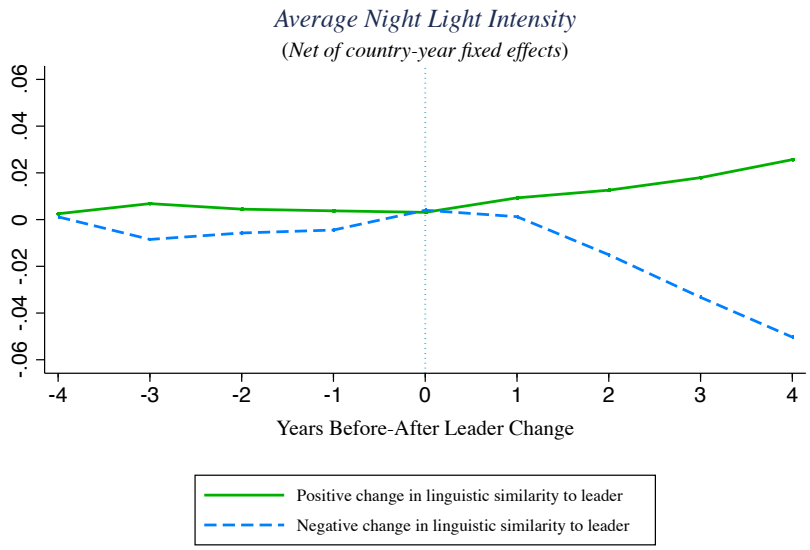
Linguistic Similarity of Leaders and Language Groups

My independent variable of interest is a measure of bilateral linguistic similarity between each country-language group partition and the ethnolinguistic identity of the country's national leader. Because the computerized lexicostatistical method requires a word list for each language of interest, I am limited to working with languages that have lists made available by the ASJP research team. Of the 227 language groups in the full set of partitions I match 163 in the benchmark regression (72%), failing the rest either because the leader's birth language list is unavailable or the partition language list is unavailable. Furthermore, 11 out of the 102 leaders ethnolinguistic identities lack an ASJP language list and are excluded from the analysis. I address the possibility of sample selection in Section 1.4. The result is an (unbalanced) panel of lexicostatistical similarity between partitioned language groups and their national leader for the years 1992-2013.[15] Figure 1.3 colour codes these groups.[16]

[15] See Appendix B for a complete list of included countries and language groups.

[16] The only other lexicostatistical data available for a large number of languages is from Dyen et al. (1992), which is restricted to Indo-European languages only – none of which are native to Africa.

Figure 1.5: Pre-Post Leadership Change



This figure plots the before and after effects of a change in leadership on average night light luminosity. The green solid line depicts luminosity in the 4 years leading up to a change in leadership and the 4 years following an increase in linguistic similarity. The blue dashed line depicts the same for country-language groups that experienced a decrease in similarity after a change in leadership. Average night light luminosity is the residual light variation net of country-year effects to account for different years of leadership change across countries.

1.2.5 Patterns in the Data

Table 1.1 reports descriptive statistics for the night lights and language data. For completeness, I have included a cladistic measure of similarity and a binary measure of coethnicity.[17] The mean value of lexicostatistical linguistic similarity says that country-language groups are 19.3 percent similar to their national leader on average, and the mean value of cladistic similarity implies 40.9 percent similarity. The mean value of coethnicity says that 4.7 percent of the benchmark sample is coethnic to their national leader.[18]

In Table 1.2 I preview the empirical results by splitting the sample by the median value of night lights and test for differences in average linguistic similarity. Panel A reports mean differences in the benchmark sample for all three similarity measures. Take, for example, the mean difference in lexicostatistical similarity: language groups who emit night light above the median value are on average 10.4 percent more similar to their national leader than those below the median value. This difference is highly significant, with a reported p-value of 0.000. The same pattern is true irrespective of the measure of linguistic similarity. These findings are consistent with my proposed hypothesis of ethnolinguistic favoritism, where language groups are better off the more similar

[17] I use the term coethnicity to be consistent with the literature, but a better name would be coethnolinguists since I define coethnicity equal to one when a leader's ethnolinguistic identity is the same as a partitioned *language* group.

[18] Table B6 reports a complete set of descriptive statistics used throughout this analysis.

Table 1.1: Descriptive Statistics

	Obs.	Mean	Std. Dev.	Min	Max
$\ln(0.01 + \text{night lights})$	6,610	-3.496	1.423	-4.605	1.515
Lexicostatistical similarity	6,610	0.193	0.230	0.000	1.000
Cladistic similarity	6,610	0.409	0.330	0.000	1.000
Coethnicity	6,610	0.047	0.212	0.000	1.000

This table reports descriptive statistics for the main variables of interest used in the benchmark empirical analysis of partitioned language groups in Africa. The unit of observation is a language group l that resides in country c in year t. See Appendix B for a description of the data and sources.

Table 1.2: Means of Linguistic Similarity Above-Below Median Night Lights

	Observations	Above Median Luminosity	Below Median Luminosity	Difference
Panel A: Full Sample				
Lexicostatistical similarity	6,610	0.245	0.142	0.104***
		(0.005)	(0.003)	(0.006)
Cladistic similarity	6,610	0.478	0.341	0.137***
		(0.006)	(0.005)	(0.008)
Coethnicity	6,610	0.078	0.016	0.062***
		(0.005)	(0.002)	(0.005)
Panel B: Non-Coethnic Regions Only				
Lexicostatistical similarity	6,298	0.182	0.125	0.057***
		(0.003)	(0.002)	(0.004)
Cladistic similarity	6,298	0.435	0.325	0.110***
		(0.006)	(0.005)	(0.008)

This table reports differences in means for various measures of linguistic similarity. Language groups are separated by the median value of night lights into "above" and "below" groups for each sample. The full sample consists of 6,610 observations and the non-coethnic subsample consists of 6,298 observations. Standard errors are reported in parentheses.

they are to their national leader.

Panel B repeats this exercise in all non-coethnic sample observations. As stated in the introduction, if relative groups differences matter outside of coethnic relationships, then the data should tell me that similarity matters among non-coethnics. This is exactly what I find: the average similarity among non-coethnic language groups above and below the median night lights value is significantly different than zero. While I reserve more conclusive statements for the regression analysis, this suggests that linguistic similarity provides significant variation that is unobservable in the conventional binary framework. Together these results show that night lights and linguistic similarity are positively related, or that on average a language group is increasingly better off the more linguistically similar they are to the birth language of their national leader. The significant pairwise correlation of 0.30 between light intensity and lexicostatistical similarity is also suggestive of this positive relationship (correlation not shown here).

I also plot average luminosity before and after a leadership change in Figure 1.5, separating groups who experience an increase in lexicostatistical similarity from those that experience a decrease. I construct a "treatment" time scale that takes a value of 0 in the year of a leadership change, and plot the residual light variation net of country-year effects to account for different years of leadership changes. I plot these data for the 4 years leading up to a change and the 4 years following. It is reassuring for identification that there is little observed change in night light activity in the years leading up to a change in leadership. Yet shortly after a leadership change there is a noticeable increase in night lights in regions that experienced an increase in linguistic similarity to the leader (solid green line), and a large drop in average night lights in regions that experienced a decrease in similarity (dashed blue line). Hence, Figure 1.5 is a clean visualization of favoritism across linguistic lines.[19]

1.3 Empirical Model

The main objective of this empirical analysis is to test the hypothesis that a language group that is linguistically similar to the ethnolinguistic identity of the national leader will be better off than a group whose language is relatively more distant. To do this I use a triple difference-in-differences estimator:

$$y_{c,l,t} = \gamma_{c,l} + \lambda_{c,t} + \theta_{l,t} + x'_{c,l,t}\,\Phi + \beta LS_{c,l,t-1} + \varepsilon_{c,l,t}. \tag{1.1}$$

The dependent variable $y_{c,l,t}$ is the night lights measure of economic activity for language group l in country c in year t. As the dependent variable I follow the literature and take the aforementioned log transformation of night lights such that $y_{c,l,t} \equiv \ln(0.01 + NightLights_{c,l,t})$.

[19] The number of observations used to calculate the average night lights in either group varies by years. The nature of the data presents two challenges in constructing a standard treatment time scale. First, in some instances there is more than one leadership change in the shown 8-year interval. Second, and in consequence of the first point, two leadership changes over the 8-year interval do not always result in consistent positive or negative changes of similarity.

$LS_{c,l,t-1}$, the variable of interest, measures the linguistic similarity between language group l in country c and the ethnolinguistic identity of country c's political leader in year $t-1$. I lag linguistic similarity because of an expected delay between the decision to allocate public funds to a region and the actual allocation of those goods (Hodler and Raschky, 2014), and an expected delay between the actual allocation of public funds and the resulting regional increase in night light production.

$X_{c,l,t}$ is a vector of controls including the (logged) average of population density for each country-language, and the (logged) geodesic distance between language group l and the language group associated with the leader of country c.[20] I also include a variety of geographic endowment controls in $x_{c,l,t}$: two indicator variables for the presence of oil and diamond reserves in both the leader and language group regions, as well as the absolute difference in elevation, ruggedness, precipitation, average temperature and the caloric suitability index (agricultural quality). These additional controls account for the possibility that national projects that are beneficial to the leader's region because of a particular geographic characteristic might also benefit other regions of similar character.[21] $\gamma_{c,l}$ are country-language group fixed effects, $\lambda_{c,t}$ are country-year fixed effects and $\theta_{l,t}$ are language-year fixed effects.[22] In all specifications I adjust standard errors for clustering in country-language groups.[23]

1.3.1 Identification of Linguistic Similarity

In order to identify the effect of linguistic similarity it is necessary that the placement of national borders are not the result of local economic conditions or any factor that reflects the well-being of a language group. Indeed, national borders are a historical by-product of the Scramble for Africa. The use of straight lines prevailed when drawing borders in Africa because the Berlin Conference of 1884-85 legitimized claims of colonial sovereignty without pre-existing territorial occupation, rendering knowledge of pre-colonial boundaries inconsequential (Englebert et al., 2002). The result was a reluctance by colonialists to respect traditional boundaries when drawing borders (Herbst, 2000). Evidence of this is still seen today, where group partitions do not correlate with geography and natural resources (Michalopoulos and Papaioannou, 2016) and nearly 80% of all African borders follow lines of latitude and longitude – an amount larger than any other continent in the world (Alesina et al., 2011).[24]

[20] Population density data comes from the Gridded Population of the World. Because population density data is only available in 5-year intervals (i.e., 1990, 1995, 2000, 2005 and 2010), I assume the density to be constant throughout the unobserved intermediate years.

[21] See Appendix B for more details on data definitions and sources.

[22] In my benchmark sample $\gamma_{c,l}$ represents 355 fixed effects, $\lambda_{c,t}$ represents 691 fixed effects and $\theta_{l,t}$ represents 3044 fixed effects.

[23] Given that the benchmark sample has only 35 countries, I choose not to adjust standard errors for two-dimensional clustering within language groups and countries (Cameron et al., 2011). While the benchmark results are qualitatively similar when two-way clustering, I follow Kezdi's (2004) rule of thumb that at least 50 clusters are needed for accurate inference.

[24] See Englebert et al. (2002) and Michalopoulos and Papaioannou (2014, 2016) for a detailed discussion on the arbitrary design of African borders.

It is the arbitrary design of African political borders that forms the basis of my identification strategy. The ethnolinguistic identity of a national leader varies by country, so group partitioning generates exogenous within-group variation in terms of that group's linguistic similarity to their leader. This strategy is similar to Michalopoulos and Papaioannou (2014), though a key difference is that I construct a panel of partitioned groups rather than a cross-section, so the relative similarity within a partitioned group also varies over time as new leaders come to power. This is instrumental to identification: by including the three sets of fixed effects discussed in the previous section, I absorb all the variation in the data with the exception of time-variation at the country-language group level. $\gamma_{c,l}$ and $\lambda_{c,t}$ respectively difference out time-invariant country-group trends and country-time trends that are differentially affecting the same group on each side of the border. The inclusion of $\theta_{l,t}$ only allows for within-group time-variation that comes from changes in leadership. Hence, with my set-up in equation (1.1), I am estimating the effect of linguistic similarity off of changes in the incoming leader's ethnolinguistic identity.

In my benchmark sample this variation comes from 35 leadership changes: the within transformation of $\theta_{l,t}$ implies that a leadership change in one country varies the mean similarity of a partition in that country and all other fragments of that partition in neighbouring countries. In other words, the relative similarity within a partitioned group varies with a leadership change on either side of the border. This amounts to 485 unique relative similarities observed between 1992-2013 in my data.

1.4 Benchmark Results

Table 1.3 reports nine different estimates: three versions of equation (1.1) for each of the three linguistic similarity measures. For each measure of similarity, I report estimates (i) without any covariates (columns 1-3), (ii) estimates that control for log population density and the logged geodesic distance between each partitioned group and the corresponding leader's group (columns 4-6), and (iii) the full set of covariates I outlined in Section 1.3 (columns 7-9). Hereafter I will refer to columns 7-9 as my benchmark specification.[25]

Consistent with my hypothesis of ethnolinguistic favoritism, all nine coefficients are positive and my preferred measure of lexicostatistical similarity is always statistically significant. Because variation is coming from changes in the ethnic identity of a leader, the interpretation of these findings is that a group's well-being is increasing in their ethnic similarity to the leader. To give economic meaning to these estimates, consider the benchmark estimate of lexicostatistical similarity in column 7. Using the rule of thumb that the estimated elasticity of GDP per capita with respect to night lights is 0.3 (Henderson et al., 2012), the point estimate of 0.305 implies that a standard deviation increase in linguistic similarity (23 percent change) yields a 2.1 percent increase in regional GDP per capita, an economically significant effect.[26]

[25] See Table B9 for various other combinations of fixed effects specifications.

[26] The percentage change in GDP per capita $\approx$ percentage change in night lights $\times 0.3 = (\beta \times \Delta LS_{c,l,t-j}) \times 0.3 =$

Table 1.3: Benchmark Regressions Using Various Measures of Linguistic Similarity

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.244**			0.297**			0.305***		
	(0.112)			(0.120)			(0.116)		
Cladistic similarity$_{t-1}$		0.221**			0.219**			0.185*	
		(0.104)			(0.102)			(0.103)	
Coethnic$_{t-1}$			0.130			0.139			0.168*
			(0.099)			(0.098)			(0.094)
Geographic controls	No	No	No	No	No	No	Yes	Yes	Yes
Distance & population density	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355	355	355	355
Countries	35	35	35	35	35	35	35	35	35
Language groups	163	163	163	163	163	163	163	163	163
Adjusted R^2	0.925	0.925	0.925	0.925	0.925	0.925	0.926	0.925	0.925
Observations	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610

This table reports benchmark estimates associating each measure of linguistic similarity with night light luminosity for the years $t = 1992 - 2013$. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t - 1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Distance & population density measure the log distance between each country-language group and the leader's ethnolinguistic group, and the log population density of a country-language group, respectively. The geographic controls include the absolute difference in elevation, ruggedness, precipitation, temperature and the caloric suitability index between leader and country-language group regions, in addition to two dummy variables indicating if both region contains diamond and oil deposits. Standard errors are clustered at the country-language group level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

I also provide estimates for cladistic similarity and coethnicity to see how these alternative measures compare to lexicostatistical similarity. For my benchmark estimates both coefficients are positive and statistically significant, albeit only at the 10 percent level. Notice that in all iterations of equation (1.1), the magnitude and precision of the estimate is monotonically increasing in the measured continuity of linguistic similarity. This suggests that the observable variation among non-coethnic groups assists in identifying patterns of ethnic favoritism in Africa, and thus speaks the virtue of the lexicostatistical measure.

In Table 1.4 I report estimates from a series of horse race regressions. With these estimates I show that the lexicostatistical measure is better at identifying patterns of favoritism than the alternative measures of similarity. In columns 1-4, I report estimates for all possible pairings of the three measures of similarity. Because all three measures of similarity are highly correlated with each other, and for coethnic observations are equivalent, the effect of lexicostatistical and cladistic similarity are estimated off of the additional variation these measures provide among non-coethnics. In all pairings the additional lexicostatistical variation is estimated to be statistically significant, despite the fact that the effect of coethnicity is not identifiable in these regressions. In column 3, cladistic similarity outperforms coethnicity in magnitude and precision, reaffirming the value of the addition variation it provides over a coethnic indicator, but is not estimated to be significantly different than zero.

To disentangle the effect of coethnicity from the benefits of similarity among non-coethnics, I define non-coethnic lexicostatistical similarity as $(1-\text{coethnic}_{t-1})\times$ lexicostatistical similarity, and equivalently for non-coethnic cladistic similarity. In other words, these non-coethnic similarity measures are equal to zero when the observed language group is coethnic to their national leader, and otherwise equivalent to the respective measure of similarity. Combined with the coethnic measure, I can exploit the same variation I identify off of in columns 2 and 3 but load the effect of coethnicity onto the coethnic dummy variable.

Because it is intuitive that a leader is more inclined to favor her coethnics, I expect to see a strong significant effect of coethnicity beyond the effect found among non-coethnic groups. Indeed, column 5 indicates that coethnics are most favored with an estimated increase of 0.260 in average night light luminosity. While there is still an observable benefit from similarity among non-coethnics, the magnitude of the effect is roughly one quarter the size of the coethnic effect on average. With a sample mean of 0.146, non-coethnic lexicostatistical similarity yields an average increase of 0.069 ($= 0.146 \times 0.473$) in night light luminosity.[27]

I repeat this exercise with non-coethnic cladistic similarity and report the estimates in column (6). Once again I find the corresponding estimate for cladistic similarity from column (3) but can now identify the effect of coethnicity. The estimated coefficient for coethnicity is quite similar to the coethnic effect found in column (5), only now the additional variation coming from the

$0.305 \times 0.230 \times 0.3 = 2.1\%$, assuming that $\ln(0.01 + \text{nightLights}_{c,l,t}) \approx \ln(\text{nightLights}_{c,l,t})$.

[27] By these estimates the threshold value of non-coethnic similarity is 0.550, above which would imply non-coethnics are better off than coethnics. The likelihood of measurement error in linguistic similarity implies this is a rather "fuzzy" threshold, and with only 2 percent of the benchmark sample above this threshold I find this result to be reassuring.

Table 1.4: Horse Race Regressions: Contrasting the Different Measures of Linguistic Similarity

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$						
	(1)	(2)	(3)	(4)	(5)	(6)
Lexicostatistical similarity$_{t-1}$	0.345**	0.473**		0.591**		
	(0.165)	(0.227)		(0.291)		
Cladistic similarity$_{t-1}$	-0.046		0.151	-0.102		
	(0.146)		(0.125)	(0.150)		
Coethnic$_{t-1}$		-0.213	0.080	-0.249	0.260**	0.230**
		(0.202)	(0.114)	(0.211)	(0.106)	(0.110)
Non-coethnic lexicostatistical similarity$_{t-1}$					0.473**	
					(0.227)	
Non-coethnic cladistic similarity$_{t-1}$						0.151
						(0.125)
Geographic controls	Yes	Yes	Yes	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355
Countries	35	35	35	35	35	35
Language groups	163	163	163	163	163	163
Adjusted R^2	0.926	0.926	0.925	0.926	0.926	0.925
Observations	6,610	6,610	6,610	6,610	6,610	6,610

This table reports horse race regressions comparing each measure of linguistic similarity. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Non-coethnic lexicostatistical similarity and Non-coethnic cladistic similarity are constructed by interacting a dummy variable for non-coethnicity with Lexicostatistical similarity and Cladistic similarity, respectively. All control variables are described in Table 1.3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p $<$ 0.10, ** p $<$ 0.05, *** p $<$ 0.01.

cladistic measure is not enough to identify the effect of similarity among non-coethnic groups.

Taken together the results of Table 1.3 and Table 1.4 indicate that favoritism is most prominent among coethnics but also to a lesser extent among non-coethnics. These results also indicate that a continuous measure of lexicostatistical similarity provides valuable information that is not observable with a coethnic indicator variable. For the remainder of this section I proceed to test the robustness of the benchmark lexicostatistical estimate.

Anticipatory Effects

In this section I run of a series of tests of the identifying assumptions underlying my benchmark estimates. Column (1) of Table 1.5 reproduces the benchmark estimate of lexicostatistical similarity for comparison. In column (2) I show that the lagged measure of lexicostatistical similarity is not essential to my findings; contemporaneous lexicostatistical similarity is estimated to be positive and significant at the 5 percent level.

In column (3) I report an estimate of lexicostatistical similarity measured in period $t + 1$. In this specification I'm estimating the effect of linguistic similarity off of the change in an incoming leader's ethnolinguistic group in the period before that leader comes to power. Should there be any pre-trends in the incoming leader's group, then this lead measure of lexicostatistical similarity should be estimated significantly different than zero. I find no evidence of a pre-trend, which is reassuring for identification that the common trends assumption is satisfied. In column (4)-(6) I report estimates from horse race regressions between lead, contemporaneous and lagged lexicostatistical similarity. Again I find no evidence of a pre-trend in the lead variable. Together these findings confirm there are no anticipatory changes in night lights preceding a change in leadership, an observation consistent with Figure 1.5. Column (6) also indicates that lagged lexicostatistical similarity is a better predictor of favoritism than contemporaneous similarity, a finding that supports my decision to lag lexicostatistical similarity.

Next I re-estimate equation (1.1) with a lagged dependent variable. Identification rests on the assumption that leaders are not endogenously elected because of the economic success of their ethnolinguistic group prior to an election. I find no evidence of this as indicated by column (7) and (8). Lexicostatistical similarity is estimated to be positive and significant at the 5 percent level, albeit with a reduced magnitude. Hence, these results are reassuring that my benchmark estimates are not an outcome of any pre-transition changes in economic activity in a leader's ethnolinguistic group.

Migration

One additional concern with my identification strategy is cross-border migration. Suppose individuals who live near the border become coethnics of the neighboring country's leader. These individuals may choose to migrate in response to this spatial disequilibrium of similarity. While the cultural affinity of partitioned groups might ease the migration process, Oucho (2006) points

Table 1.5: Testing for Anticipatory Effects: Estimates Using Leads and Lags

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Lexicostatistical similarity$_{t-1}$	0.305***			0.299***		0.249**	0.170**	0.131**
	(0.116)			(0.110)		(0.101)	(0.067)	(0.062)
Lexicostatistical similarity$_{t}$		0.495**			0.406**	0.242		0.214
		(0.204)			(0.205)	(0.183)		(0.134)
Lexicostatistical similarity$_{t+1}$			0.170	0.134	0.059	0.067		0.021
			(0.117)	(0.107)	(0.096)	(0.096)		(0.070)
Night lights$_{t-1}$							0.521***	0.506***
							(0.050)	(0.055)
Geographic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355	355	351
Countries	35	35	35	35	35	35	35	35
Language groups	163	163	163	163	163	163	163	161
Adjusted R^2	0.926	0.926	0.930	0.930	0.930	0.930	0.947	0.950
Observations	6,610	6,474	6,121	6,121	6,084	6,084	6,315	5,785

This table reports a series of tests for anticipatory effects in the benchmark estimates. Average night light intensity is measured in language group l of country c in year t, and Lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the national leader and is measured on the unit interval. The same log transformation of the dependent variable is used for the lagged value of night lights, i.e., $\ln(0.01 + NightLights_{c,l,t-1})$. All control variables are described in Table 1.3. Standard errors are clustered at the country-language group level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

out that migration restrictions throughout sub-Saharan Africa make this unlikely in a formal capacity, so this might only be an issue among undocumented migrants. Not only do undocumented migrants make up a small percentage of total migrants but those that do migrate tend to do so to trade and are temporary by definition (Oucho, 2006). To corroborate this anecdotal evidence, I also regress log population density on linguistic similarity in period $t-1$ and report the estimates in Table 1.6. If people are in fact migrating in response to leadership changes, I should observe corresponding changes in population density. These estimates also account for the possibility of within-country migration. In all specifications, the various measures of similarity are insignificant. Overall these estimates imply that changes in night lights within a partitioned group cannot be explained by movements of people to regions that are similar to the leader in terms of ethnolinguistic identity.

Sample Selection

My inability to observe the lexicostatistical similarity of the 64 language groups without an ASJP language list raises the question whether these unobserved groups are systematically different than those in my benchmark sample. To address this concern I test for mean differences in key observables and report these differences in Table 1.7.

First I show that there is no difference in the average night light luminosity between in- and out-of-sample partitioned language groups. I also show that there is no difference between the cladistic similarity of in- and out-of-sample groups. These two results are reassuring that both sets of partitioned groups are comparable in terms of economic activity and proximity to their leader.

To the contrary, I show that in-sample groups reside in countries that are, on average, more democratic, more competitive politically, have more constraints on the executive, and are more open and competitive in the recruitment of executives. Should there be an in-sample selection bias, these institutional mean differences suggest that my estimates would be biased towards zero, given the evidence that a well-functioning democracy mitigates the extent of ethnic favoritism (Burgess et al., 2015) and regional favoritism (Hodler and Raschky, 2014)

Robustness Checks

I also show that the results are robust to a variety of specifications and estimators. I report and discuss each robustness check in Appendix B. In particular, I show that the results are similar when:

- I reproduce my benchmark estimates with additional controls for malaria and land suitability for agriculture. Because these data are only available at a $0.5^\circ \times 0.5^\circ$ spatial resolution (approx. 111 km $\times$ 111 km), I exclude them from my benchmark estimates to avoid losing observations where a pixel is larger than a language group partition (Table B10).

Table 1.6: Test for Migration Following Leadership Changes

Dependent Variable: ln(Population Density$_{c,l,t}$)			
	(1)	(2)	(3)
Lexicostatistical similarity$_{t-1}$	0.001		
	(0.019)		
Cladistic similarity$_{t-1}$		-0.027	
		(0.031)	
Coethnic$_{t-1}$			0.010
			(0.016)
Country-language fixed effects	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes
Clusters	355	355	355
Countries	35	35	35
Language groups	163	163	163
Adjusted R^2	0.999	0.999	0.999
Observations	6,610	6,610	6,610

This table reports estimates associating population density with linguistics similarity as a test for changes in population density following a change in a leader's ethnolinguistic identity. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Standard errors are clustered at the country-language group level and are reported in parentheses. * p $<$ 0.10, ** p $<$ 0.05, *** p $<$ 0.01.

- I check that measurement error coming from ambiguous assignment of a leader's ethnolinguistic identity does not explain my benchmark results, particularly the finding that favoritism exists among non-coethnics (Table B11).
- I reproduce my benchmark estimates on a balanced panel of 84 ethnolinguistic groups partitioned across 23 countries (Table B12).
- I re-estimate equation (1.1) and weight the estimates by the Ethnologue population of each language group (Table B13). The idea here is to correct for possible heteroskedasticity: the measure of night light intensity is an average within each country-language group, and it is likely to have more variance in places where the population is small.
- I also provide estimates with two alternative transformations of the night lights data to show that my benchmark lexicostatistical estimate is not an outcome of the aforementioned log transformation (Table B14).

Table 1.7: Selection into Lexicostatistical Language Lists

		Partitioned Language Groups		
	Observations	Benchmark Sample Mean	Out of Sample Mean	Difference
$\ln(0.01 + \text{night lights})$	11,869	-3.487	-3.505	0.018
		(0.018)	(0.022)	(0.028)
Cladistic similarity	11,869	0.276	0.272	0.004
		(0.004)	(0.004)	(0.005)
Level of democracy (Polity2)	11,822	0.677	0.319	0.358***
		(0.059)	(0.062)	(0.086)
Political competition	10,854	6.180	5.940	0.239***
		(0.032)	(0.033)	(0.046)
Executive constraints	10,854	3.634	3.368	0.266***
		(0.022)	(0.022)	(0.032)
Openness of executive recruitment	10,854	2.756	2.556	0.200***
		(0.024)	(0.028)	(0.036)
Competitiveness of executive recruitment	10,854	1.283	1.208	0.075***
		(0.014)	(0.015)	(0.021)

This table tests for selection into the available language lists in the ASJP database. The full sample of partitioned language groups are separated by those that I observe in my benchmark dataset and those that I do not because of missing ASJP language lists. Standard errors are reported in parentheses.

1.4.1 What Drives Favoritism?

In this section I test for heterogeneity across a variety of potential channels to better understand what drives favoritism. In Table 1.8 I study the dynamics of my benchmark findings by account for the possibility that the extent of favoritism is a function of the time a leader has held office. In column 1, I report estimates of an augmented equation (1.1) that includes an interaction between linguistic similarity and a count of the years a leader has held office. The interaction term enters positive and statistically significant, indicating that favoritism is an increasing function of the years a leader has held office.

I also construct a set of indicator variables at 5-year intervals to explore the non-linearities further. Column 2 reports these estimates. All coefficients are positive and the magnitude of effect is increasing in the length of tenure, however there is no significant effect associated with the first five leaders of leadership. Taken together, Table 1.8 indicates that the extent of ethnolinguistic favoritism is an increasing function of a leader's incumbency. In a continent where multi-decade presidencies are not uncommon (e.g., Jose Eduardo dos Santos in Angola or Robert Mugabe in Zimbabwe), it should come as no surprise that favoritism is so rampant.

I also check for heterogeneous effects across seven other measures: the level of democracy (Padró i Miquel, 2007; Burgess et al., 2015), language group Ethnologue population shares (Fran-

cois et al., 2015), distance to the capital from a group's centroid (Michalopoulos and Papaioannou, 2014), distance to the nearest coast from a group's centroid (Nunn, 2008; Nunn and Wantchekon, 2011), presence of an oil reserve and diamond mine within the territory of the country-language group (Jensen and Wantchekon, 2004). Table 1.9 reports these estimates.

The analysis reveals little evidence of heterogeneity. One explanation for a lack of heterogeneity is that these different channels are only relevant in some countries and do not generalize to the 35-country sample I use here. Another possible explanation is that the rich set of fixed effects in each regression absorb much of the important variation. For example, in column (1), I find that democracy has a mitigating effect on the extent of observed favoritism, but this effect is not statistically significant. While the intuition is consistent with Burgess et al. (2015), the lack of precision likely comes from the fact that country-year fixed effects account for the level effect of democracy, and the residual variation is not significant enough to identify any meaningful effect. A similar explanation applies to the remaining variables, where country-language fixed effects absorb the level effect for each because of the time invariance of these group-level measures.

However, there is some evidence of heterogeneity in terms of a diamond mine being present within a country-language group. The negative coefficient implies favoritism is less prevalent in regions where diamond mines exist. On interpretation is that the presence of diamonds creates wealth, and the resulting development may reduce the material importance of patronage to the region. Yet the lack of heterogeneity in oil reserves does not corroborate this story, so I leave a more concrete analysis of why diamond mines might constrain favoritism to future research.

1.5 How Is Patronage Distributed?

In this section I develop a within-group set-up similar to the previous section using individual-level data from the Demographic and Health Survey (DHS). Exploiting the same source of variation with different data serves as an additional robustness check of my benchmark analysis. The DHS data also allows me to explore how patronage is distributed. In particular, I use data on an individual's location and ethnolinguistic identity to construct two measures of lexicostatistical similarity: locational and individual similarity. I define individual similarity as the lexicostatistical similarity of the leader to each respondent's ethnolinguistic identity. To assign a locational language I use the Ethnologue language map and individual location coordinates to determine the language group associated with an individual's location. I define locational similarity as the lexicostatistical similarity of a leader to the respondent's locational language. Because these measures do not always coincide, I can jointly estimate both channels to determine the relative importance of being similar to the leader versus living in a location with an attached identity similar to the leader.

Table 1.8: The Dynamics of Ethnolinguistic Favoritism

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$		
	(1)	(2)
Lexicostatistical similarity$_{t-1}$	0.072	
	(0.160)	
Lexicostatistical similarity$_{t-1}$	0.027*	
$\times$ Years in office$_{t-1}$	(0.016)	
Lexicostatistical similarity$_{t-1}$		0.118
$\times$ 1(Years in office$_{t-1} \leq 5$)		(0.129)
Lexicostatistical similarity$_{t-1}$		0.325*
$\times$ 1($5 <$ Years in office$_{t-1} \leq 10$)		(0.170)
Lexicostatistical similarity$_{t-1}$		0.561***
$\times$ 1($10 <$ Years in office$_{t-1} \leq 15$)		(0.197)
Lexicostatistical similarity$_{t-1}$		0.555**
$\times$ 1($15 <$ Years in office$_{t-1} \leq 20$)		(0.233)
Lexicostatistical similarity$_{t-1}$		0.689**
$\times$ 1($20 <$ Years in office$_{t-1}$)		(0.347)
Geographic controls	Yes	Yes
Distance & population density	Yes	Yes
Language-year fixed effects	Yes	Yes
Country-language fixed effects	Yes	Yes
Country-year fixed effects	Yes	Yes
Clusters	355	355
Countries	35	35
Language groups	163	163
Adjusted R^2	0.926	0.926
Observations	6,610	6,610

This table reports estimates of the dynamics of ethnolinguistic favoritism. The unit of observation is a language group l in country c in the specified year. Average night light intensity is measured in language group l of country c in year t, and Lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the ethnolinguistic identity of the national leader. Current years in office is a count variable of the years the incumbent leader has been in power, and total years in office measures the total years the incumbent leader will remain in power. Quartile measures relate to current years in office. All control variables are described in Table 1.3. Standard errors are clustered at the country-language group level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.9: Benchmark Regressions with Heterogeneous Effects

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$						
	(1)	(2)	(3)	(4)	(5)	(6)
Lexicostatistical similarity$_{t-1}$	0.298**	0.407**	0.379**	0.327*	0.305***	0.397***
	(0.127)	(0.161)	(0.174)	(0.190)	(0.116)	(0.135)
Lexicostatistical similarity$_{t-1}$	-0.005					
$\times$ Democracy$_{t-1}$	(0.020)					
Lexicostatistical similarity$_{t-1}$		-0.610				
$\times$ Population share		(0.533)				
Lexicostatistical similarity$_{t-1}$			-0.000			
$\times$ Distance to the capital			(0.000)			
Lexicostatistical similarity$_{t-1}$				-0.000		
$\times$ Distance to the coast				(0.000)		
Lexicostatistical similarity$_{t-1}$					0.232	
$\times$ Oil reserve					(1.140)	
Lexicostatistical similarity$_{t-1}$						-0.336*
$\times$ Diamond mine						(0.190)
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355
Countries	35	35	35	35	35	35
Language groups	163	163	163	163	163	163
Adjusted R^2	0.927	0.926	0.926	0.926	0.926	0.926
Observations	6,540	6,610	6,610	6,610	6,610	6,610

This table reports a series of tests for heterogeneous effects in the benchmark estimates. Average night light intensity is measured in language group l of country c in year t, and lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the national leader and is measured on the unit interval. All control variables are described in Table 1.3. Democracy is the polity2 score of democracy for the country in which a group resides, geodesic distances are measured in kilometres from a group's centroid to the capital city and the nearest coast, oil reserve and diamond mine represent indicators variables at the group level. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

1.5.1 DHS Individual-Level Data

I collect data from the Demographic and Health Surveys (DHS) for 13 African countries.[28] For each country I pool both the male and female samples for each wave, and when separately provided, I merge the wealth index dataset for that year. To replicate the same variation I use in my benchmark estimates, I choose DHS countries and survey waves as follows:

(1) I identify all DHS country-waves that include latitude and longitude coordinates for each survey cluster as well as information on a respondent's home language and/or ethnic identity.

(2) I identify all language groups that are partitioned across contiguous country pairs in the DHS database that also possess the necessary information noted in (1).

(3) For each partitioned language group identified in (2) I keep those that possess at least 2 consecutive surveys from the same set of DHS waves.

Next I project the latitude and longitude coordinates for each survey cluster onto the Ethnologue language map and back out the language group associated with that location.[29] I assign this language as the locational language for that cluster and construct a measure of locational similarity as the lexicostatistical similarity of that region to the incumbent leader.

To measure individual similarity I use data on the language a respondent speaks at home, and when not available data on their ethnicity. I describe the mapping between ethnicity and language in detail in Appendix B. I construct a measure of individual similarity as the lexicostatistical similarity between the home language of an individual and the ethnolinguistic identity of their national leader. To be consistent with my benchmark model, I measure locational and individual linguistic similarity to the national leader in year $t-1$.

The result is 33 DHS country-waves, 13 countries and 11 country pairs, with 20 partitioned language groups. Having at least 2 consecutive survey waves for each partitioned group allows for a set-up similar to my benchmark model, where within-group time variation comes from leadership changes across waves. One important difference from my benchmark set-up is that for 3 countries I only observe a single partitioned language group, meaning that country-location-language fixed effects are not applicable in this context.

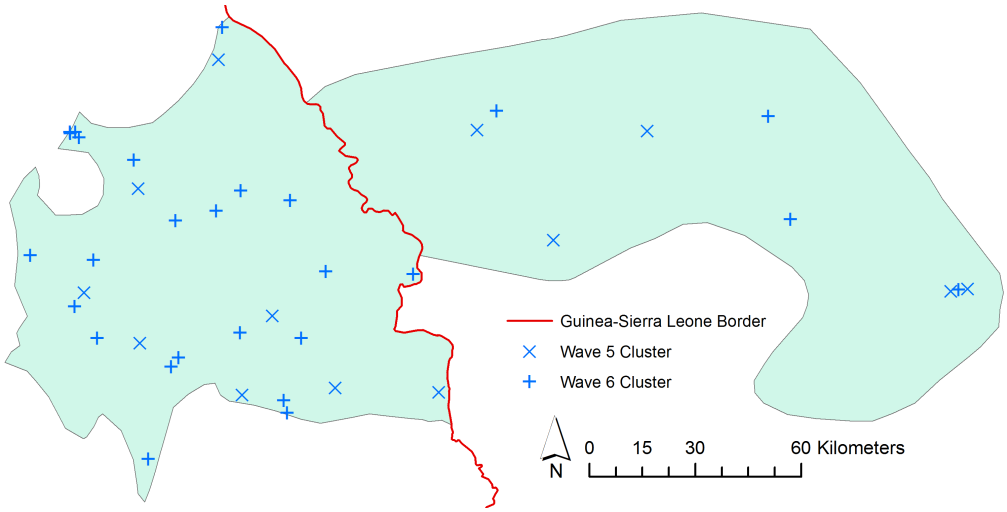
Among the 56,455 respondents for whom I successfully match both locational and individual languages, I find that 55.9 percent reside in their ethnolinguistic homeland.[30] This finding corroborates the implicit assumption made in the regional-level analysis that the majority of a language region's inhabitants are native to that region. At the same time, having 44.1 percent of respondents

[28] See Appendix B for a list of countries and a detailed discussion of all DHS data.

[29] In instances of overlapping language groups, I assign the largest group in terms of population

[30] Nunn and Wantchekon (2011) also find that 55 percent of respondents in the 2005 Afrobarometer reside in their ethnolinguistic homeland. The consistency across datasets is quite remarkable since only 7 out of the 13 countries used in this paper overlap with the Afrobarometer data in Nunn and Wantchekon (2011).

Figure 1.6: DHS Clusters Across Waves in the Kuranko Language Group Partition



This figure documents the spatial distribution of DHS enumeration clusters in the partitioned Kuranko language group in Sierra Leone (west of the border) and Guinea (east of the border). Variation between individual and locational lexicostatistical similarity comes from the 40 percent of respondents who identify with an ethnolinguistic group different than the Kuranko.

be non-native to their location allows me to exploit variation in individual and location similarity to separately estimate the two effects off of leadership changes.[31]

Consider, as an example, the Kuranko language group partitioned across Guinea and Sierra Leone. Figure 1.6 depicts the spatial distribution of Kuranko survey clusters by wave. For each survey respondent living in one of these clusters I assign the Kuranko language as their locational language, despite the fact that only 60.1 percent of respondents report Kuranko as their ethnolinguistic identity. Among the remaining 39.9 percent of respondents in the Kuranko region there are 9 other reported ethnolinguistic identities. Take the 117 Sierra Leoneans living in the Kuranko region who identify as Themne – the ethnicity/language of president Ernest Bai Koroma. The inclusion of individual similarity allows me to ask if Themne respondents benefit from coethnicity – and similarity more generally – irrespective of where they live.

1.5.2 Locational and Individual Similarity Estimates

I test the general importance of locational and individual similarity vis-à-vis changes in the DHS wealth index – a composite measure of cumulative living conditions for a household. The index is constructed using data on a household ownership of assets (e.g., television, refrigerator, telephone, etc.) and access to public resources (e.g., water, electricity, sanitation facility, etc.). Since variation in linguistic similarity comes from leadership changes, a positive estimate for either measure implies better access to public resources and more acquired assets because of an individual's

[31] The use of non-natives in this way is methodologically similar to Nunn and Wantchekon (2011) and Michalopoulos et al. (2016), who also use variation within non-native Africans to disentangle regional effects from individual-level effects.

similarity across the significant dimension.

In every specification I include country-wave fixed effects, locational language-wave fixed effects and individual language-wave fixed effects. As previously mentioned I do not include country-language fixed effects because in some instances I only observe a single language for a country. Unlike estimating equation (1.1), I include individual language-wave fixed effects because 45 percent of respondents' home language is different than their locational language.

I report these estimates in Table 1.10. In column 1 the estimate for lexicostatistical locational similarity is positive and significant at the 1 percent level. This point estimate of 0.540 is equivalent to 0.35 of a standard deviation increase in the wealth index. In column 2 I report the estimate for individual similarity. While the estimate has the expected positive sign, the coefficient is not precisely estimated. This suggests that changes in the individual-level wealth index are coming from transfers based on the ethnic identity of a region. Indeed, when I run a horse race between the two, I find that locational similarity is significantly different than zero while individual similarity remains insignificant. [32]

Overall, these estimates indicate that favoritism operates through regional transfers, which suggests that favoritism is beneficial to all inhabits of a region regardless of their background. This finding is consistent with the evidence that Kenyan leaders invest twice as much in roads (Burgess et al., 2015), and disproportionately target school construction in their coethnic districts (Kramon and Posner, 2016). In a case study of Congo-Brazzaville, Franck and Rainer (2012) similarly find that ethnic divisions impact the patterns of regional school construction. However, this case study also points to anecdotal evidence of the individual-level channel, where coethnic individuals benefit from preferential access to education and civil servant jobs irrespective of where they live. Kramon and Posner (2016) similarity posit the existence of this preferential access channel. To the contrary, I find that an individual's similarity to her leader does not afford her any luxuries beyond the location effect.

Finally, to show that the locational mechanism is not only driven by the coethnic effect, I separately estimate locational coethnicity and non-coethnic locational similarity. I do this in the same way I did in the regional-level analysis: I define non-coethnic locational similarity as $(1 - \text{coethnicity}) \times \text{locational similarity}$. Column 4 of Table 1.10 reports this estimate. I find that both the coethnic and non-coethnic effect are positive and strongly significant. These estimates suggest that the average level of non-coethnic locational similarity (0.164) yields an increase of 0.122 ($= 0.164 \times 0.742$) in the wealth index – roughly one fourth the coethnic effect.

1.6 Discussion: Coalition Building

The results of this paper indicate that ethnic favoritism is widespread throughout Africa, and that patronage is distributed to both the ethnic region of the leader and to related but non-coethnic regions. But what mechanism drives these regional transfers? Why might we expect to see fa-

[32] See Appendix B for the unconditional estimates.

Table 1.10: Individual-Level Regressions: Locational and Individual Similarity

Dependent Variable: DHS Wealth Index				
	(1)	(2)	(3)	(4)
Locational similarity$_{t-1}$	0.540***		0.541***	
	(0.128)		(0.128)	
Individual similarity$_{t-1}$		0.239	0.240	
		(0.216)	(0.216)	
Locational coethnicity$_{t-1}$				0.501***
				(0.133)
Non-coethnic locational similarity$_{t-1}$				0.742***
				(0.148)
Individual controls	Yes	Yes	Yes	Yes
Distance controls	Yes	Yes	Yes	Yes
Country-wave fixed effects	Yes	Yes	Yes	Yes
Locational language-wave fixed effects	Yes	Yes	Yes	Yes
Individual language-wave fixed effects	Yes	Yes	Yes	Yes
Clusters	88	88	88	88
Countries	13	13	13	13
Language groups	20	20	20	20
Adjusted R^2	0.605	0.605	0.605	0.605
Observations	56,455	56,455	56,455	56,455

This table reports estimates that test for favoritism outside of coethnic language partitions. The unit of observation is an individual. The rural indicator is equal to 1 if a respondent lives in a rural location. The individual set of control variables include age, age squared, rural indicator variable, a gender indicator variable and an indicator for respondents living in the capital city. Distance to the coast and border are in kilometers. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * p $<$ 0.10, ** p $<$ 0.05, *** p $<$ 0.01.

voritism outside of the leader's ethnic region?

I offer an explanation that relates to the literature on coalition building. Central to this literature is the idea that leaders respond to political instability by co-opting elites from outside of their ethnic group into high-level government positions to pacify unrest and to maintain control of the state (Joseph, 1987; Arriola, 2009; Francois et al., 2015). The fact that similar but not identical ethnic regions benefits from patronage suggests that ethnicity is more than just a marker of identity: similarity may capture affinity between related non-coethnic groups. It is intuitive that the "closeness" of a group to the leader would predict their share in the governing coalition for reasons related to trust (Habyarimana et al., 2009), reduced costs of coordination (Miguel and Gugerty,

2005), clientelistic networks (Wantchekon, 2003) and more. Because leaders share power with ethnic groups other than their own to make credible their promise of patronage (Arriola, 2009), any evidence that leaders appoint closely related groups is an indication that coalition building is one mechanism underlying this paper's findings.

An insightful paper by Francois et al. (2015) provides theoretical and empirical support for the claim that ethnic group representation in the governing coalition is proportional to a group's share of the national population. The logic of this theory runs contrary to ethnic favoritism: their proposed mechanism underlying coalition building is group size. Yet these authors still find that a leader's ethnic group receives a premium in government appointments over and above the effect of group size. While it is beyond the scope of this paper to take a stance on the relative importance of these channels, what is important is that they are not mutually exclusive to each other.

To shed light on this interesting area of research I document that the similarity of an ethnic group to the leader correlates with an ethnic group's representation in government *conditional* on group size. I use yearly data from Francois et al. (2015) on the share of an ethnic group's representation in the governing coalition for 15 African countries between 1992 and 2004.[33] The majority of Ethnologue groups are defined as Others in Francois et al.'s (2015) data, which severely limits the observable number of group partitions. Consequently, it is not possible to use the same source of within-group variation employed elsewhere in this paper. Instead I use an identical set-up to Francois et al. (2015), but augment their empirical model with an indicator variable for similar but not identical ethnic groups:

$$y_{c,e,t} = \alpha \, \text{coethnic}_{c,e,t} + \beta \, \text{similar}_{c,e,t} + f\left(\text{groupsize}_{c,e}\right) + \delta_c + \gamma_t + \epsilon_{c,e,t}. \tag{1.2}$$

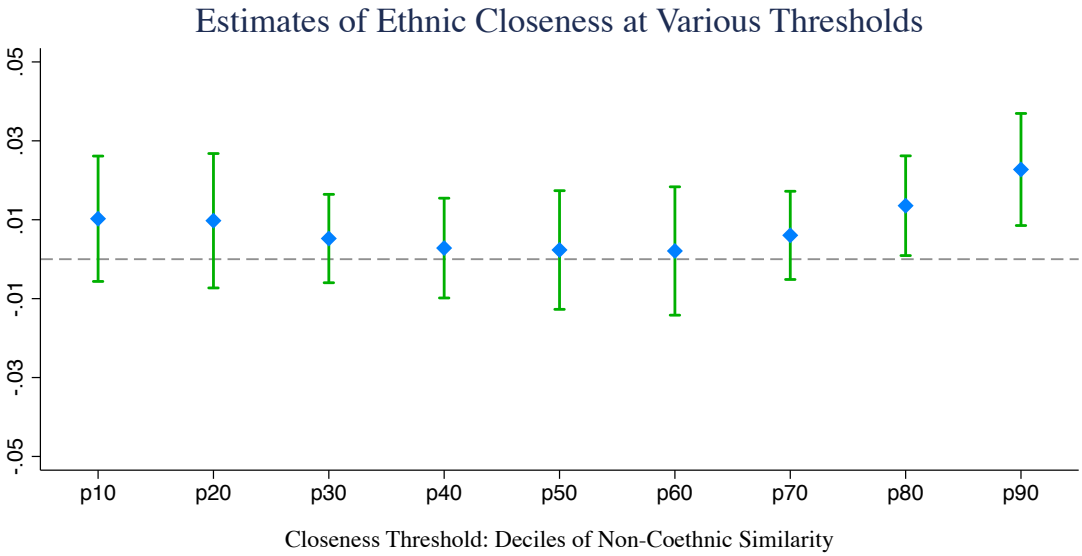
The outcome variable $y_{c,e,t}$ is ethnic group e's share of cabinet positions in country c in year t. In addition to the usual $\text{coethnic}_{c,e,t}$ indicator variable, I include an indicator equal to one when a non-coethnic group's linguistic similarity is greater than a defined threshold of "closeness" (i.e., $\text{similar}_{c,e,t}$). Francois et al. (2015) find that $\text{groupsize}_{c,e}$ – the population share of ethnic group e in country c – is concave in its relationship with a group's share of cabinet positions, so I include $\text{groupsize}_{c,e}$ and its polynomial in all regressions. δ_c and γ_t capture unobserved time-invariant country effects and time trends. I follow Francois et al. (2015) and cluster standard errors at the country level.

I exploit a range of thresholds to let the data inform me of the relevant threshold of closeness. My preferred approach is to split the distribution of linguistic similarity for non-coethnics into deciles. I assign the threshold for closeness as any non-coethnic observation with similarity to the leader that is equal to or greater than a defined decile of the distribution. Hence, these thresholds are cumulative, where $\text{similar}_{c,e,t}$ is equal to one when a non-coethnic group's linguistic similarity is equal to or greater than the decile cut-off.[34]

[33] See Appendix B for details on the construction of this dataset.

[34] In Appendix B, I replicate Table III from Francois et al. (2015) and include the more general specification of lexicostatistical similarity in place of coethnicity.

Figure 1.7: Ethnic Favoritism and Coalition Building



This figure plots point estimates of the $\text{similar}_{c,e,t}$ indicator variable in equation (1.2) at nine different closeness thresholds. Thresholds are set according to deciles of the distribution for non-coethnic similarity. These thresholds are cumulative, where $\text{similar}_{c,e,t}$ is equal to one when a non-coethnic group's linguistic similarity is equal to or greater than the decile threshold. Each estimate reflects the additional share of cabinet positions a similar non-coethnic group receives relative to non-similar non-coethnics. Intervals reflect 99% confidence levels.

I plot the point estimates of β in Figure 1.7 for various thresholds, where the intervals reflect 99 percent confidence levels. The figure clearly documents non-coethnic favoritism in coalition building, at least among the stricter definitions of closeness. Because I include $\text{coethnic}_{c,e,t}$ and $\text{similar}_{c,e,t}$ in each regression, the estimate of β reflects the additional share of cabinet positions a similar non-coethnic group receives *relative* to other non-coethnics that do not satisfy the threshold level of similarity.

To assess the economic meaning of these estimates, I can compare the difference in a group's predicted outcome conditional on *mean* group size, after turning β on and off. Let the closeness threshold be the most stringent threshold at the 90^{th} percentile of the distribution. I find that a non-coethnic group's share in the governing coalition jumps from 5.1 to 7.4 percent when β is included – a 45 percent increase.[35] The resulting share is almost 2 percentage points larger than the sample average share of 5.6 percent.

But how similar are these "close" groups? Consider the Gbe ethnolinguistic family, where three of the most widely spoken languages include Fon, Ewe and Gen. For the three possible pairings of these languages, the average lexicostatistical similarity is 46.8 percent. The mean similarity among non-coethnic groups in the 90^{th} percentile of the distribution is 45.7 percent. This

[35]The point estimate for $\text{similar}_{c,e,t} = 0.023$, while the point estimates for group size and its polynomial are 1.225 and -1.795. For the mean non-coethnic group, the effect of group size is $5.1 = 1.225 \times 0.052 - 1.795 \times 0.007$, and 7.4 when adding the $\text{similar}_{c,e,t}$ premium.

suggests that leaders appoint elites from outside of their immediate ethnic group that are part of the same family cluster. In other words, the affinity that similarity captures is reflective of the shared ancestry in a group's larger ethnic network.

Overall, these findings suggest that leaders are inclined to make ethnicity-based decisions when appointing ministers from outside their own ethnic group. While the estimates of equation (1.2) cannot necessarily be taken as causal, they are informative of the mechanism through which public resources are allocated to non-coethnic regions. The tendency of ministers to redirect funds to their coethnics explains why non-coethnic groups with representation in government receive patronage (Arriola, 2009).

1.7 Concluding Remarks

Ethnic favoritism is often thought to be endemic to African politics, yet the empirical basis for this claim is largely founded on single-country case studies. In this paper, I document evidence that ethnic favoritism is widespread throughout Africa using data for 35 sub-Saharan countries. I also introduce a novel measure of linguistic similarity that contributes to this literature in three ways: (i) it better predicts patterns of ethnic favoritism with added variation in measured similarity, (ii) the continuity of the measure enables the study of favoritism among groups that are not coethnic to the leader, and (iii) it informs our understanding of a new mechanism related to the ethnic affinity between similar but non-coethnic groups. This deepens our understanding of the *extent* of favoritism – evidence of favoritism among non-coethnics normally goes undetected when using a coethnic dummy variable. I also show that patronage tends to be distributed at a regional level rather than as targeted transfers towards individuals. I interpret these results through the lens of coalition building and find that ethnicity is one of the guiding principles behind high-level government appointments.

These observations inform policy in a number of new ways. In particular, my findings are informative of both the extent of favoritism and where it is expected to take place. This can be used for many purposes, one of which is to enhance monitoring of foreign aid. There is a growing body of evidence that links the misuse of foreign aid to ethnic patronage in Africa (Briggs, 2014; Jablonski, 2014). Greater oversight is achieved through a deeper understanding of where patronage is expected to flow. My findings suggest aid donors should not only worry about patronage directed toward the leader's ethnic group but also toward other related groups. The benefits of oversight are evident when comparing the non-interference aid policy of China with conditional transfers from the World Bank. Dreher et al. (2015) find little evidence that World Bank aid is used for patronage purposes in contrast to the evidence that China's non-interference policy engenders resource allocation across ethnic lines rather than on a basis of need.

More generally, my findings speak to the value of nation-building policies that promote diversity in a Pan-Africanist tradition. Tanzania is a good example of a country that has stressed a sense of unity and shared history in its national policies. One nation-building tool of this type

that is particularly relevant to this paper is Tanzania's national language policy (Miguel, 2004). In the mid-1960s, the Tanzanian government changed the official language of the country to Swahili. The extent to which Swahili is found in other countries and commonly used as a lingua franca speaks to the ethnic neutrality of the language. Within only a few years of its implementation, the official status granted to Swahili was described as a "linguistic revolution" for its ability to help shape a national consciousness that runs contrary to ethnic identity (Harries, 1969, p. 277). The Tanzania example is a model to be replicated elsewhere, given the evidence that ethnic favoritism is so widespread throughout Africa. This is not to imply that national language policies are the only means to pacify existing ethnic tensions: the lesson here is that national policies must be designed to engender acceptance of diversity through unity. For example, education is an effective way to build a national culture that actively values diversity and differences in experience and background.

Lastly, the findings of this paper call for future work. The evidence that favoritism is not simply a coethnic phenomenon demands a deeper understanding of what it means to be "close" to the ruling ethnic group. The taxonomy of linguistic and ethnic clusters provide an opportunity to study this notion of closeness in the same vein as Desmet et al. (2012). Linking the extent of favoritism to the impact it has on ethnic inequality is an important next step in this line of research. The Tanzania example also suggests ethnic favoritism is not an unavoidable consequence of a country's high level of diversity, an observation that is consistent with the literature on ethnic inequality. Why then do we observe favoritism in some countries and not others? Geographic segregation is linked to ethnic favoritism in Africa (Ejdemyr et al., 2014), while geographic endowments are linked to ethnic inequality (Alesina et al., 2016), which suggests an answer to this question lies at the intersection of these two areas of research.

Chapter 2

Population Relatedness and Cross-Country Idea Flows

2.1 Introduction

Recent research documents a link between the ancestral relationship of two countries and their current difference in income (Spolaore and Wacziarg, 2009, 2013a). This link is interpreted as reflecting an indirect causal effect: income gaps are smaller between related populations because they are more likely to communicate and adopt similar ideas. By this interpretation the probability of an idea flowing between two countries is the indirect causal link, and this probability is smaller in more distant relationships. At the same time, dissimilarity could theoretically provide incentive for idea flows if a wider spectrum of non-overlapping traits increase the likelihood of two populations having complementary ideas. This notion of a diversity-driven incentive for idea flows suggests a possible counterbalancing force to the lower probability of communication when two countries are ancestrally distant. This hypothesis is similar in spirit to the theory of Ashraf and Galor (2013) and Ashraf et al. (2014, 2015), who document empirical evidence of opposing forces of relatedness on income per capita within a country.

The purpose of this research is then two-fold: can the degree of relatedness between countries explain the cross-country flow of ideas, and if so, do we observe interplay between two opposing forces of population relatedness? I find that the degree of relatedness can explain these flows, and that measuring relatedness across linguistic and genetic dimensions yields robust empirical evidence of two opposing forces on the cross-country flow of ideas.

I follow the literature in using genetic distance as a measure of population relatedness, and use data on book translations to capture bilateral flows of information. Consistent with the suggested mechanism of Spolaore and Wacziarg (2009), I find that unconditional genetic distance negatively correlates with book translations. However, after isolating variation in population differences specific to language and geography, the residual variation indicates book translations are increasing in genetic distance. This is consistent with the proposed diversity-driven communication incen-

tive. The observed negative correlation of unconditional genetic distance captures a cost specific to language that I account for when linguistic distance is jointly estimated. The result is two measures of relatedness capturing opposing forces: language differences impose a cost on idea flows while population differences yield a communication incentive.

I also find that linguistic distance reflects a stronger relationship with book translations (in absolute terms) than genetic distance. This evidence reconciles my findings with the intuition of Spolaore and Wacziarg (2009), where unconditional genetic distance is expected to exhibit a negative relationship with book translations. One reason genetic distance has received so much attention in the literature is because it captures a variety of intergenerationally transmitted traits – notably language (Spolaore and Wacziarg, 2015). As a summary measure it is then intuitive that unconditional genetic distance absorbs the overarching cost of language differences and negatively correlates with book translations. The stable and positive relationship between genetic distance and book translations is observable only after separating out linguistic and geographic variation from genetic distance.

Conversely the estimated coefficient for linguistic distance is stable in significance and magnitude with or without conditioning on genetic distance. The stability of the estimate has two implications: that linguistic distance is not a latent measure of genetic distance and that language constrains the flow of ideas.

Book translations are an appealing measure of idea flows since by definition they require a bilateral exchange in terms of language and often location. This feature lends itself to bilateral comparisons of population relatedness. The bilateral nature of book translations also makes them suitable for study within a gravity model, the empirical model most frequently used in studies of genetic and linguistic differences. Book translations are also a recognized measure of international idea flows (Abramitzky and Sin, 2014), and satisfy the necessary properties of an idea because they are non-rival and disembodied.[1]

But more than this a book can convey an idea or tell a story that reflects a different set of values and cultural norms, or draw analogy as a form of commentary about the state of society. The breadth of ideas that a book can capture make translations an appealing measure of idea flows because of a book's capacity to influence human behaviour. Rodrik (2014) argues ideas are not only relevant as technical innovations, but also shape our preferences and how we think the world works, and at times "can unlock what otherwise might seem like the iron grip of vested interests" (p. 194). This behavioural influence of an idea resonates with a long history of books and their influence over society.[2]

Using an empirical gravity model of translation flows, I estimate that a one standard deviation increase in linguistic distance yields 12 percent fewer book translations. This distance is roughly

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

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

equivalent to how much more Romanian differs from English than German. This result is highly significant and robust to a range of controls, including per capita income and population, political rights, and other covariates measuring bilateral differences in geography and colonial history. Estimates also hold in magnitude and significance when disaggregating the translation data by idea type, as well as including numerous measures of human capital and bilateral trade between country pairs. For the benchmark estimate I flexibly control for country, time and language fixed effects, and later show that the core linguistic distance result holds even when accounting for unobserved country-pair effects. The stability of the linguistic distance estimate in significance and magnitude shows little evidence of a selection bias driving this benchmark result.

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

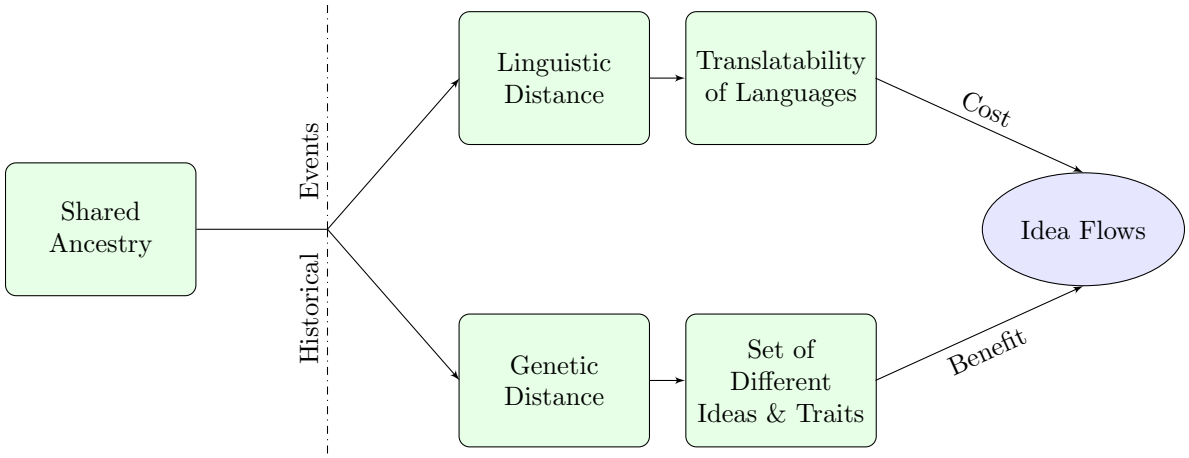
To interpret these findings I argue relatedness affects both the costs and benefits of social interactions overall. On the one hand, the translatability of two languages is a crucial determinant of communication between countries. Translatability reflects the cost of capturing the original message of a book in a translation – I use a computerized lexicostatistical measure of distance to capture this cost. On the other hand, there is little overlap of traits and cultural norms when ancestral relationships are distant. This wider spectrum of separate traits increase the likelihood of complementary ideas and therefore an incentive for communication exists when groups have more to learn from each other, possibly driving the positive relationship between genetic distance and book translations.

It is also intuitive that we observe this positive relationship through genealogical differences. By construction, genetic distance reflects a molecular clock that maps the time since two populations shared a common ancestor (discussed further in section 2.2.2). Although linguistic distance reflects the time since two populations shared a common language, this does not necessarily capture the same historical relationship. For example, the Magyar invaded Hungary in the ninth century, imposing their Uralic language on the conquered. To this day Hungarians exhibit a gene distribution similar to the rest of Central Europe, but continue to speak a Uralic language unlike the Latin-based languages of their neighbours (Cavalli-Sforza, 2000). A similar pattern exists in many regions of the transatlantic slave trade, where colonizers imposed their own language on the colonized with little genetic mixing (Phillipson, 1992). Such historical events create a wedge between the co-evolution of language and genetics. In this sense genetic distance is a better summary measure of the time since two populations separated, and thus the extent of dissimilar ideas, beliefs and cultural norms. Figure 2.1 depicts this argument schematically.

The principal contribution of this paper is the evidence that population relatedness confers both social costs and benefits on the flow of ideas. Ashraf and Galor (2013) document a trade-off between the costs and benefits of diversity on productivity *within* a country.[3] While I am

[3] Ashraf et al. (2014, 2015) also document a trade-off between the costs and benefits of genetic diversity on night light

Figure 2.1: Opposing Forces of Population Relatedness on the Flow of Ideas



concerned with idea flows, the intuition gleaned from Ashraf and Galor (2013) suggests that the interplay between these opposing forces of relatedness also exists *between* countries. I document evidence of this between-country link, which offers a deeper understanding of Spolaore and Wacziarg's (2009) proposed mechanism: summary population differences exhibit a negative relationship with the diffusion of ideas, with the caveat that this negative relationship operates along linguistic lines and that a concomitant incentive for idea flows exists along distant ancestral lineages.

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

The use of book translation data as a measure of idea flows also places this research in proximity to a recent paper by Abramitzky and Sin (2014), who document the repressive nature of communist institutions on the inflow of Western books in the Soviet Union prior to its collapse. I

luminosity at the national level and subnational national, respectively.

account for the institutional environment of a country with a measure of political rights, but find that the coefficient estimates of linguistic and genetic distance are unaltered in significance and magnitude conditional on the extent of political rights. This suggests the influence of the institutional environment is only part of the story, and that the degree of relatedness is a contributing factor to the bilateral flow of book translations across countries.

The rest of this paper is structured as follows. Section 2.2 describes the translation data and details the measurement strategy for the linguistic and genetic distance data. I outline the econometric model in section 2.3 and document the benchmark empirical finding. Section 2.4 tests the robustness of the benchmark result and section 2.5 extends the analysis by disaggregating the book translation data by idea type. Section 2.6 concludes.

2.2 Data

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

2.2.1 Measuring Language Distance

Measurement of linguistic differences is difficult because languages can differ in a variety of ways, including vocabulary, pronunciation, grammar, syntax, phonetics and more. To overcome this challenge I isolate variation specific to vocabulary differences with a computerized lexicostatistical approach. As a percentage estimate of cognate words in a language pair, the lexicostatistical method measures how the passage of time diminishes the lexical similarity between two languages.[4] The recent splitting of two languages from a parent language implies a large number of shared cognates or a small linguistic distance.

The computerized lexicostatistical method was developed as part of the *Automatic Similarity Judgement Program* (ASJP), a project run by linguists at the Max Planck Institute for Evolutionary Anthropology.[5] To begin a list of 40 universal words (i.e., implied meanings) are compiled for each language to compare the lexical similarity of any language pair. Swadesh (1952) introduced this idea of comparing a universal list of words. When a word is universal across world languages, its implied meaning, and therefore any estimate of linguistic distance, is independent of culture and geography.[6]

Each word in each language is 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. Words are then transcribed ac-

[4]Two words are said to be cognate if they exist in separate languages but share a common linguistic origin. In other words they descend from the same historical parent language.

[5]`http://asjp.clld.org`

[6]A recent paper by Holman et al. (2009) shows that the 40-item list employed here yields results at least as good as those of the commonly used 100-item list propo1sed by Swadesh (1955). Specifically, the 40-item list is a refined version of the 100-item list, deduced from rigorous testing for word stability across all languages.

cording to pronunciation before language distances are estimated. For example, the French word for *you* is *vous*, and is encoded using ASJPcode as *vu* to reflect its pronunciation.

I then take these word lists and run the Levenshtein distance algorithm to calculate the minimum number of edits needed to translate a word from one language into another. To correct for the fact that longer words demand more edits, each distance is divided by the length of the translated word. This normalization procedure yields a percentage estimate of dissimilarity. Averaging across the 40-item list of words for each language pair yields an average lexicostatistical distance between two languages.

Following Wichmann et al. (2010), a second normalization procedure is used to adjust for the accidental similarity of two languages. This normalization accounts for similar ordering and frequency of characters that are the result of chance and independent of a word's meaning.

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. To account for this I divide each distance by the maximum distance, the result of which is a measure of lexicostatistical linguistic distance measured across the unit interval. See Appendix A for a formal definition of this lexicostatistical language distance.[7]

Existing studies typically use data from Fearon (2003), a measure of cladistic language distance. The cladistic measure is a ratio of shared branches between two languages on the language tree, and distance is calculated as one minus this ratio. This measure is widespread in its use because cladistic distance is easily computed for any language pair, since language trees exist for virtually all known world language families (Lewis, 2009).[8]

The main advantange of the lexicostatistical measure is that it is a continuous measure of linguistic distance and thus provides significant variation in distance estimates. The cladistic approach is a coarse measure of distance because data dispersion is limited to 15 unique values, the maximum number of language family classifications on the *Ethnologue* langauge tree. Conversely I match 72 unique lexicostatistical distances to my sample of book translations. This additional variation gives me a distribution of distances among language pairs that would otherwise exhibit no variation in cladistic distance. Dickens (2016a) uses this lexicostatistical measure to approximate the similarity of African ethnolinguistic groups to their national leaders, and empirically verifies the added precision it yields in estimation of ethnic favoritism over and above the cladistic measure.

[7] A number of other economists have used the data from Dyen et al. (1992) as an estimate of lexicostatistical linguistic distances for the Indo-European family (Desmet et al., 2005, 2009, 2011; Ku and Zussman, 2010; Spolaore and Wacziarg, 2009, 2014, 2016). But this data is restricted to 84 Indo-European languages and does not employ the computerized approach used here. The non-computerized approach calls for a trained linguist to work with each possible cognate one by one to make judgement of cognation among them. Such an approach relies on subjectively determined cognates, and is extremely labour intensive, thus inhibiting the number of language comparisons possible. Greenberg (1956) formally introduced this approach, and Dyen et al. (1992) provide a detailed discussion of the procedure.

[8] See Appendix A for a formal discussion of the cladistic approach.

2.2.2 Measuring Genetic Distance

Data on cross-country genetic distance comes from Spolaore and Wacziarg (2009), who collected their data from Cavalli-Sforza et al. (1994). To quantify genetic distance, Cavalli-Sforza et al. (1994) collected data on population allele frequencies specific to a set of selectively neutral genes. An allele is one of many different forms the same gene can assume, where different phenotypic traits (observable characteristics) develop out of different allele sets. Genes were chosen that are known to be selectively neutral to ensure genetic variation across populations is the result of genetic drift. The random nature of drift makes genetic differences simply a function of time. Comparing the distribution of neutral allele frequencies effectively measures the time since two populations shared a common ancestor, so genetic distance becomes a molecular clock.[9]

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

When constructing a measure of genetic distance at the country level it is problematic that many countries contain multiple ethnic sub-groups. To correct for this I adopt a genetic distance measure weighted by the population share of each sub-group in a country.[10] This measures the expected genetic distance between a randomly selected individual from each country. All reported empirical results use this weighted measure versus one that calculates the distance between the dominant population in each country. However, the two measures are highly correlated and the core empirical result is robust to this alternative measure.[11]

2.2.3 Book Translations as Idea Flows

I use a bibliographic database on book translations from around the world as a measure of international idea flows. While book translations are not the only means to knowledge diffusion, they are perhaps the most common form of transmitting written ideas between countries, making them a useful and quantifiable measure of idea flows. Abramitzky and Sin (2014) also use book translations as a measure international idea flows to study the diffusion of information under different institutional regimes.

One appealing feature of book translations is the breadth of ideas they capture. Rodrik (2014) argues that the influence of an idea extends beyond the technical innovation of a patent citation,

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

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

[11] Because the F_{st} genetic distance data has become more common in the economics literature, I have foregone some details in how the measure is constructed for the sake of brevity. I direct interested readers to Spolaore and Wacziarg (2009, p. 480) for a thoughtful and detailed discussion of the data.

and that economists should recognize a broader scope of ideas that may also influence human behaviour more generally.

Acknowledging the power and influence of books is at the heart of my measurement strategy. Technical ideas are not excluded from the book translation data, but unlike patent citations, these data also capture innovative activity outside the scope of technology. Innovative ideas include new ways of thinking about societal norms, and how individual interests are defined and pursued. Leighton and López (2013) argue the rules of society shape the incentive structures that we live by, and that new ideas change these rules, which in turn provide new incentives and ultimately an influence over human behaviour.

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

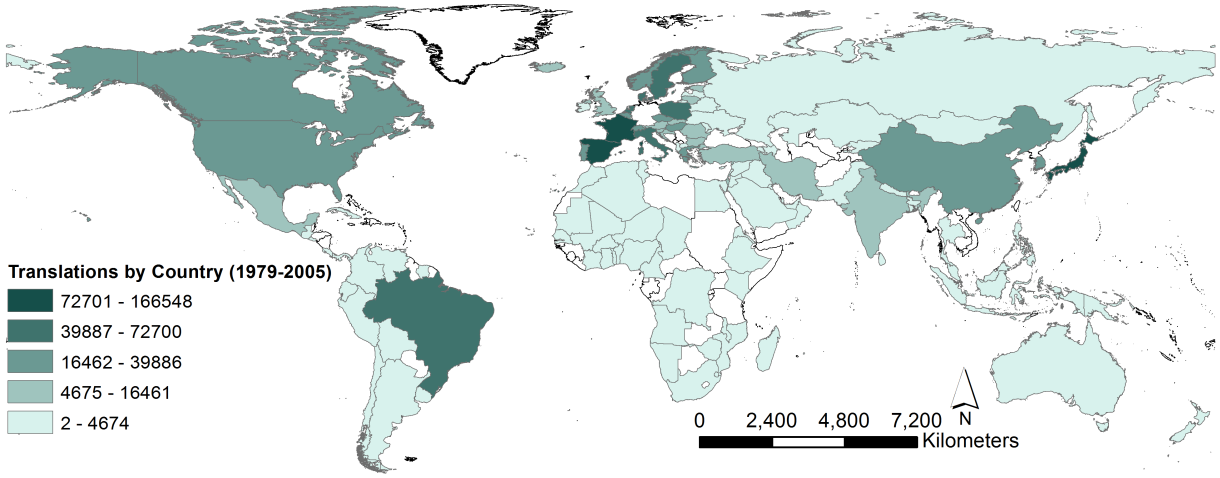
The influence of books is not only apparent in controversy, but also in their capacity to disseminate ideas. Israel (2009) argues that the transatlantic democratic revolutions of the late eighteenth century have an intellectual origin rooted in the ideas of the Enlightenment, which "persuaded much of the reading elite on either side of the Atlantic [...] that a general revolution in the principle and construction of governments is necessary" (p. 39). This wave of democratic revolutions, Israel argues, was propelled by the spread of pamphlets and books articulating these ideas. This example also speaks to the economic importance of books in the long-run. Given the recent evidence that democracy causes growth (Acemoglu et al., 2014), by extension the intellectual origin of democracy indirectly links the historical role of books to development patterns of today.

Measuring Book Translations

Translation data was collected from the Index Translationum (IT) database, an international bibliographic archive of book translations hosted by the United Nations Educational, Scientific and Cultural Organization (UNESCO). Since 1932, IT has compiled a detailed record of book translations in print form, and more recently developed an online archive containing bibliographic information of translated books in UNESCO Member States since 1979. Legal deposit legislation states that all publications of a book be submitted to the book repository of a country.[12] Records of book translations are then submitted to IT by the national book repository. The systematic nature of the data collection process is a reassurance of the data's accuracy.

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

Figure 2.2: Benchmark Sample Observations by Country (1979-2005)



I use data on 1,634,817 book translations that span 119 translating countries over the time period 1979-2005. The panel is unbalanced because I do not observe the same translating country–language pairs in each year. Figure 2.2 depicts the spatial distribution of observations for the benchmark sample of translating countries.[13] While the majority of these translations are into the dominant language of the translating country, many countries translate books into more than one language. Each bibliographic entry contains information by subject, the translating country and year, and the original and target languages of translation.[14] Table C2 lists some of the most translated authors within a random sample of countries, including the subject in which they have most commonly published.

Because the translation data does not report the country in which a book was originally published, I assign a home country to each original translation language as stated in the *Ethnologue*.[15] The benefit of this assignment rule is that it removes judgement and hence any personal bias in the data construction. The drawback is that, in some cases, multiple countries share a common official language so it is not clear the *Ethnologue's* stated home country is always the correct one. For example, it is unlikely that all translations originating in English were originally published in the United Kingdom as the *Ethnologue* assignment rule would suggest. In section 2.4.1 I perform two tests of this assignment rule: I drop the most problematic languages in terms of home country assignment and alternatively construct a home language "country" as an average of country characteristics weighted by population shares for each country in which an original language is found. In either case the benchmark result remains unaltered. Given the reassurance of this resounding evidence I proceed with the *Ethnologue* assignment rule as a consistent method of home country assignment for each original language of translation.

[13] See Table C3 in the appendix for a complete list of observations by translating country.

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

[15] The *Ethnologue* is a comprehensive database cataloguing all of the world's 7,097 known living languages.

2.3 Methodology and Empirical Results

2.3.1 Econometric Model

Given the bilateral nature of linguistic and genetic distance, I adopt an empirical model similar in spirit to the gravity equation. The basic theoretical gravity model implies that bilateral trade between two countries is a function of their economic size and a variety of costs to trade, notably geographic distance. With this in mind, I develop a similar estimation strategy that tests how linguistic and genetic distance affect bilateral translation flows. This basic relationship can be written as:

$$\begin{aligned} TRANS_{ijlt} \;=\; & \alpha_0 \;+\; \alpha_1 LINGDIST_{ijl} \;+\; \alpha_2 GENDIST_{ij} + \; \alpha_3 GEODIST_{ij} \\ & +\; \mathbf{X}'_{ij}\mathbf{\Gamma} \;+\; \mathbf{X}'_{it}\mathbf{\Omega} \;+\; \mathbf{X}'_{jt}\mathbf{\Phi} \;+\; \gamma_i \;+\; \gamma_j \;+\; \gamma_l \;+\; \gamma_t \;+\; \varepsilon_{ijlt} \end{aligned} \tag{2.1}$$

where i indexes the translating country, j the original country (and language), l the target language and t the time period.[16] The dependent variable $TRANS$ measures log translations, $LINGDIST$ and $GENDIST$ measure linguistic and genetic distance, and $GEODIST$ measures the geographic distance in 10,000 kilometre units. $\mathbf{X}_{ij}$ is a vector of bilateral time-invariant measures of geography and colonial relations, while $\mathbf{X}_{it}$ and $\mathbf{X}_{jt}$ include time-varying measures of income, population and political rights in country i and j. A set of fixed effects is also included in each regression, capturing unobserved country effects, time effects and idiosyncratic target language effects. In all regressions I estimate robust standard errors clustered at the country-pair level.

Reverse causality is not a problem here unless one believes that the amount a country translates has an effect on the distance between two languages or the genetic make-up of those populations. The implausibility of this comes from the fact that group differences are historically determined long before the time period of interest to this study. To account for the possibility of an omitted variable bias I include a rich set of covariates and fixed effects. I control for a variety of standard measures of trade costs and a host of time-varying measures believed to influence book translations. I also find that the benchmark estimates are robust to measures of time-varying bilateral trade and human capital. Country fixed effects absorb any unobserved country-specific factors while time effects absorb year-specific shocks, and robustness checks confirm that unobserved, time-invariant factors specific to bilateral country pairs do not drive the empirical findings.

2.3.2 Unconditional Benchmark Results

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

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

Table 2.1: Summary Statistics for Distance Measures

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

Note: 42,817 observations for correlations and all summary statistics.

Table 2.1 provides summary statistics for these measures. As expected, all three distance measures are positively correlated with each other. Linguistic distance exhibits a positive correlation of 0.34 with genetic distance and 0.24 with geographic distance, and genetic and geographic distance exhibit a positive correlation of 0.45. Book translations exhibit negative correlation with each distance measure, including -0.11, -0.12 and -0.12 with linguistic, genetic and geographic distance. No pairwise correlation between two distance measures is very large in magnitude, and yet all exhibit pairwise negative correlation with book translations, suggesting that each distance measure may have a significant and differential effect on book translations.

Table 2.2 reports the unconditional estimates of the regression analysis. Columns (1) through (3) indicate that all measure of distance negatively correlate with book translations when separately estimated. The systematic negative relationship between each unconditional distance and book translations is intuitively consistent with the interpretation of Spolaore and Wacziarg (2009). The estimate for linguistic distance is not only statistically significant but also economically meaningful: a standard deviation increase in linguistic distance implies an 18.6 percent decrease in book translations ($\exp(0.12 \times -1.72) - 1 = 0.186$). The magnitude of this effect is greater than that of genetic distance, where the estimate in column (2) implies a standard deviation increase in genetic distance yields a 11.9 percent decrease in book translations ($\exp(0.04 \times -3.17) - 1 = 0.119$). At the bottom of the table I provide standardized "beta" coefficients to simplify relative comparisons. Column (1) indicates that a one standard deviation increase in linguistic distance accounts

Table 2.2: Unconditional Benchmark Regressions

	Dependent variable: Log translations			
	(1)	(2)	(3)	(4)
Linguistic distance	-1.72***			-1.52***
	(0.26)			(0.26)
Genetic distance		-3.17***		2.87***
		(1.05)		(0.96)
Geographic distance			-0.86***	-0.85***
			(0.14)	(0.15)
Translating Language FE	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	42,817	42,817	42,817	42,817
Adjusted R^2	0.27	0.26	0.27	0.28
Country pair clusters	2112	2112	2112	2112
Standardized Coefficients (%)				
Linguistic distance	-12.7	-	-	-11.2
Genetic distance	-	-8.3	-	7.5

Country-pair clustered robust standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

for 12.7 percent of a standard deviation decrease in log book translations, whereas column (2) indicates a one standard deviation increase in genetic distance accounts for 8.3 percent of a standard deviation decrease in log book translations.

To explore the possibility of a differential effect of these two measures further I run a horse race regression and report the estimates in column (4). The coefficient estimate for linguistic distance is remarkably stable in magnitude and significance. This is reassuring that linguistic distance is not a latent measure of genetic and geographic differences, but instead an accurate measure of summary language differences. The standardized coefficient for linguistic distance falls by 1.5 percentage points from -12.7 in column (1) to -11.2 in column (4) after controlling for both genetic and geographic distance.

Conversely, after conditioning on both linguistic and geographic distance the correlative relationship between genetic distance and book translations becomes positive. Yet despite this change in sign the estimate remains statistically significant at the 1 percent level. The standardized beta coefficient implies a one standard deviation increase in genetic distance accounts for 7.5 percent of a standard deviation increase in book translations. The sensitivity of the coefficient is also suggestive that genetic distance does in fact capture summary population differences, including linguistic and geographic differences.

The influence of geographic distance is also of a notable magnitude: Norway is expected to

translate over 3 percent more books originating in neighbouring Sweden than Finland based on geographic distance alone, all else being equal. On a global scale the influence of geographic distance can become quite large.

Taken together, the unconditional estimates of Table 2.2 nicely summarize the earlier discussion of the opposing forces of relatedness on idea flows. For book translations, linguistic distance reflects a cost that hinders interactions between linguistically distinct populations so that on average a country can be expected to translate more books from a country that speaks a similar language. Genetic distance has the opposite effect; a country can be expected to translate more books written by authors of distant genealogical ancestries. Hence the trade-off between a lower probability of book translation when two countries are linguistically distant and a diversity-driven incentive for translation when two countries share distant ancestries.

2.3.3 Conditional Benchmark Results

When interpreting the core results of the previous section, one source of concern is that book translations are driven by time-varying country-specific factors unaccounted for by country fixed effects. To be sure the results of the previous section are not confounded by an omitted variable bias, I test the robustness of the distance measures to a variety of covariates.[17]

Column (1) of Table 2.3 reproduces column (4) of Table 2.2 using the benchmark sample.[18] In column (2) I report estimates that include measures of log real GDP per capita and log population for both the translating and original country. Adding these time-varying measures leaves the standardized value for linguistic distance unchanged and still maintains significance at the 1 percent level. The estimate of genetic distance also maintains statistical significance and is unchanged in magnitude. Similarly, adding political rights has no observable effect on the magnitude and significance of both distances, as shown in column (3). Given that these covariates tend to change at a slow pace this result isn't surprising because country fixed effects absorb the majority of the observable variation.

In column (4) I report estimates that include a measure of log bilateral trade. Bilateral trade flows not only measure the extent of an economic partnership, but they also capture latent determinants of trade such as existing communication networks or the extent of bilateral trust between countries (Guiso et al., 2009). In this sense it is intuitive that a country pair that trades more would also be more inclined to share ideas. Indeed, the coefficient estimate for bilateral trade indicates an influence on book translations: a 10 percent increase in bilateral trade yields a 1.2 percentage increase in book translations. This suggests that the extent of a bilateral trade relationship has a small positive but significant influence on the level of book translations for a given country pair. At the same time the standardized coefficient for linguistic distance falls 2 percentage points in magnitude when conditioning on trade, while the standardized coefficient for genetic distance exhibits a slight increase.

[17] See Appendix C for data definitions and sources.

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

Table 2.3: Conditional Benchmark Regressions

	Dependent variable: Log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	-1.55***	-1.55***	-1.54***	-1.27***	-1.24***	-1.13***
	(0.28)	(0.28)	(0.28)	(0.28)	(0.28)	(0.28)
Genetic distance	3.21***	3.19***	3.19***	3.32***	3.28***	2.40**
	(1.04)	(1.04)	(1.05)	(1.03)	(1.02)	(1.05)
Geographic distance	-0.85***	-0.85***	-0.85***	-0.57***	-0.55***	-1.37***
	(0.16)	(0.16)	(0.16)	(0.16)	(0.17)	(0.46)
Log real GDP per capita		-0.08	-0.07	-0.17**	-0.18**	-0.15**
translating country		(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
Log real GDP per capita		0.11**	0.11**	0.03	0.02	0.04
original country		(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
Log population		0.12	0.24	0.38**	0.39**	0.39**
translating country		(0.18)	(0.17)	(0.18)	(0.18)	(0.18)
Log population		0.09	0.12	0.26**	0.28**	0.30**
original country		(0.12)	(0.12)	(0.13)	(0.13)	(0.13)
Political rights			-0.21**	-0.21***	-0.21***	-0.22***
translating country			(0.09)	(0.08)	(0.08)	(0.08)
Political rights			-0.07	-0.04	-0.03	-0.04
original country			(0.08)	(0.07)	(0.07)	(0.07)
Log bilateral trade				0.12***	0.13***	0.10***
				(0.02)	(0.02)	(0.02)
$= 1$ if ever in a					-0.13	-0.11
colonial relationship					(0.14)	(0.13)
$= 1$ for contiguity						0.14*
						(0.08)
Absolute difference in latitude						-0.00
						(0.00)
Absolute difference in longitude						0.01**
						(0.00)
Translating Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	39,275	39,275	39,275	39,275	39,275	39,275
Adjusted R^2	0.28	0.28	0.28	0.28	0.28	0.28
Country pair clusters	1897	1897	1897	1897	1897	1897
Standardized Coefficients (%)						
Linguistic distance	-11.0	-11.0	-11.0	-9.0	-8.8	-8.0
Genetic distance	8.4	8.4	8.4	8.7	8.6	6.3

Country-pair clustered robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

To the contrary, when accounting for the spatial proximity of a country pair the influence of genetic distance falls by over 2 percentage points. The drop in magnitude of genetic distance is in consequence of adding longitudinal difference to the set of covariates, since genetically distant populations tend to be geographically distant as well. The positive and significant coefficient on absolute difference in longitude in column (6) is interesting because it suggests ideas tend to flow between geographically distant countries across similar latitudes. The presence of a north-south friction is consistent with the historical evidence that information tends to flow across east-west axes (Diamond, 1997; Blouin, 2014). At the same time, contiguity suggests translations are more likely between neighbouring countries. While these two results are contradictory on the surface, they are consistent with the idea of counterbalancing forces in idea flows that is at the heart of this paper. The cost of translation is lower between neighbouring countries where languages tend to be similar, and yet ideas flow between distant countries because the probability of a wider spectrum of traits and ideas provides incentive for communication.[19]

Hereafter I refer to the estimates in column (6) as my benchmark estimates. Using this specification to assess the magnitude of these estimates, first note that a standard deviation in linguistic distance reduces the dependent variable by 0.124 log units, while a standard deviation in genetic distance increases the dependent variable by 0.096 log units.[20] This says that a standard deviation increase in linguistic distance reduces book translations by 11.7 percent ($\exp(-0.124) - 1 = -0.117$), while a standard deviation increase in genetic distance increases book translations by 10.1 percent ($\exp(0.096) - 1 = 0.101$).

2.4 Robustness

2.4.1 Testing the Home Country Assumption of Book Translations

A concern of the benchmark result is that the home country of a book translation is not known. This implies that genetic distance is measured with error. As an assignment rule I use the home country of a language as indicated in the *Ethnologue* database. However, this rule does not allow for the original language of a book translation to be the same in more than one source country.

Problematic Languages

As a first test of the home country assumption I identify the most problematic languages and purge them from my benchmark dataset. Using data from the *Ethnologue* I generate a list of lan-

[19] As further evidence of these conditional benchmark estimates, I reproduce the estimates of Table 2.3 in Table C6 using a cladistic measure of linguistic distance. The opposing forces of linguistic and genetic distance hold even when using this alternative measure of linguistic distance.

[20] See Appendix Table C1 for a complete list of the benchmark sample summary statistics.

Table 2.4: Robustness Check for Problematic Original Languages of Translation

	Dependent variable: Log translations						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Excluding:	English Language Translations	Arabic Language Translations	French Language Translations	Mandarin Language Translations	Spanish Language Translations	Hindi Language Translations	All Six Language Translations
Linguistic distance	-1.48***	-1.10***	-1.04***	-1.14***	-1.12***	-1.13***	-1.39***
	(0.25)	(0.28)	(0.27)	(0.28)	(0.28)	(0.28)	(0.24)
Genetic distance	2.02**	2.41**	1.96*	2.28**	2.45**	2.40**	1.44*
	(0.88)	(1.04)	(1.03)	(1.05)	(1.08)	(1.05)	(0.81)
Geographic distance	-1.43***	-1.40***	-1.14**	-1.35***	-1.29***	-1.37***	-1.06***
	(0.38)	(0.46)	(0.45)	(0.48)	(0.47)	(0.46)	(0.40)
Benchmark controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Target Language FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Target Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	34,292	38,032	36,198	38,342	37,727	39,021	27,238
Adjusted R^2	0.33	0.29	0.29	0.28	0.28	0.28	0.36
Country pair clusters	1840	1833	1860	1880	1871	1892	1694
Standardized Coefficients (%)							
Linguistic Distance	-12.0	-7.8	-7.9	-8.1	-8.0	-8.0	-13.0
Genetic Distance	5.9	6.3	5.3	5.8	6.5	6.3	4.5

This table reports estimates on various subsamples that exclude problematic languages in terms of ambiguous home country assignment. All regressions include the benchmark set of control variables used in column (6) of Table 2.3. Country-pair clustered robust standard errors reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

guages sorted by first-language speakers. The top five languages by number of speakers include Mandarin (1,197 million), Spanish (414 million), English (335 million), Hindi (260 million) and Arabic (237 million). I also use the *Ethnologue* to generate a list of languages sorted by the number of countries in which the language is spoken. The top five languages by this definition include English (99 countries), Arabic (60 countries), French (51 countries), Mandarin (33 countries) and Spanish (31 countries). Because these languages are spoken in so many countries they are problematic in terms of assigning a single country to each language. It is no surprise these lists are almost identical. The result is a list of six potentially problematic languages: English, Arabic, French, Mandarin, Spanish and Hindi. As a test of home country assignment I re-estimate the benchmark model and systematically drop all books translated from a problematic language. Table 2.4 presents these results.

All estimates are statistically significant with the expected sign, and the majority are comparable to the benchmark estimate in terms of magnitude. The first estimate that stands out is reported in column (1), where the magnitude of the linguistic distance estimate is quite sensitive to the sample restricting English translations. This result is not surprising given that English is a lingua franca for so many speakers of different first languages around the world. The ubiquity of English as a global language is evident in the fact that only one out of four users of the language are native speakers (Crystal, 2003). Consider the comparable standardized beta of -8.0 percent for linguistic distance in the full sample. The significant increase in this standardized coefficient to -12.0 percent suggests the negative effect of linguistic distance is quite a bit stronger when abstracting from translations originating in English. This result is reassuring because it implies that the benchmark specification conservatively estimates the effect of linguistic distance when no consideration is given to the importance of English as a global lingua franca.

The other estimate that differs considerably from the benchmark estimate is reported in column (7), where I drop all translations originating in any of the six problematic languages. The standardized coefficient for linguistic distance jumps up to -13.0 percent, while the genetic distance coefficient falls to 4.5 percent. As a test of the home country assignment this result is promising because it says that, if anything, the assumption of a home country for the most problematic languages attenuates the benchmark linguistic distance estimate. While the coefficient for genetic distance drops below the magnitude of the benchmark estimate, it remains robust to a 45 percent loss of observations relative to the benchmark sample. This provides further evidence that a diversity-driven incentive for idea flows exists between countries.

In the appendix I extend this test by systematically dropping every original language of a translation with at least 100 observations in the benchmark sample. Table C7 reports these results. In every one of these 45 additional regressions, the coefficients of interest maintain the expected sign and statistical significance at standard levels of confidence. While the cut-off of 100 observations is arbitrary, the coefficient estimates converge to that of the benchmark estimates after the cut-off because the sample restrictions become so small.

Table 2.5: Conditional Benchmark Regressions with Alternative Home Country Assignment

	Dependent variable: Log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	-1.63***	-1.63***	-1.63***	-1.47***	-1.48***	-1.36***
	(0.25)	(0.25)	(0.25)	(0.25)	(0.25)	(0.26)
Genetic distance	2.56**	2.56**	2.56**	2.51**	2.47**	1.95*
	(1.04)	(1.04)	(1.04)	(1.03)	(1.04)	(1.08)
Geographic distance	-0.64***	-0.65***	-0.64***	-0.43***	-0.43***	-0.24
	(0.11)	(0.11)	(0.11)	(0.12)	(0.12)	(0.35)
Economic controls	No	Yes	Yes	Yes	Yes	Yes
Political controls	No	No	Yes	Yes	Yes	Yes
Trade controls	No	No	No	Yes	Yes	Yes
Colonial controls	No	No	No	No	Yes	Yes
Geography controls	No	No	No	No	No	Yes
Target Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Target Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	37,690	37,690	37,690	37,690	37,690	37,690
Adjusted R^2	0.29	0.29	0.29	0.29	0.29	0.30
Country-language clusters	2950	2950	2950	2950	2950	2950
Standardized Coefficients (%)						
Linguistic distance	−11.3	−11.3	−11.3	−10.2	−10.3	−9.4
Genetic Distance	6.4	6.4	6.4	6.3	6.2	4.9

This table reproduces the benchmark estimates using a sample with synthetic home "language" country assignment. Controls are constructed as a weighted average of all countries in which an original language is found. As weights I use global language population shares from the Ethnologue for each country of an original language. The set of economic controls include log real GDP per capita and log population in both countries, political controls include political rights in both countries, trade controls include the logged value of bilateral trade flows of the country pair, the colonial controls indicate if a country pair has every been in a colonial relationship, and the geography controls indicate contiguity, and a country pair's absolute difference in latitude and longitude. Robust standard errors are clustered on translating country–original language pairs and reported in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Alternative Home Country Assignment

As a second test of the home country assumption I show that the benchmark estimates are robust to an alternative assignment strategy. I construct a synthetic language home "country" as a weighted average of all countries in which an original language is found. As weights I use global language population shares from the *Ethnologue* for each country of an original language. This approach is similar in spirit to Putterman and Weil's (2010) migration matrix. For genetic and geographic distance I measure bilateral distance between the translating country and every country of an original language, and construct an average measure of distance weighted by the population share of a language in each original country.[21] I construct a weighted average of bilateral trade and the absolute difference in latitude and longitude in the same way. The weighted average of a

bilateral indicator variable becomes a percentage measure of the indicator for all possible country pairs of the original language, and non-bilateral measures including real GDP, population and political rights are a simple weighted average. Given the alternative structure of this data I replace original country fixed effects in equation (2.1) with original language fixed effects, and cluster standard errors by translating country and original language pairs instead of country pairs.

The benefit of a synthetic language home "country" over the *Ethnologue* home country assignment is that all origin countries of a language are proportionally present in covariate calculations. By doing so I relax the assumption of a home country for each language. The drawback is that I cannot construct a weighted average with complete coverage of all groups because I lack country-level data for some countries of language origin. To account for languages with particularly poor coverage I trim the bottom 5 percent of languages according to coverage.[22] The average coverage of these trimmed languages is only 56 percent and range from 0 to 90 percent. The average coverage of the remaining languages is 99 percent with a median of 100 percent. This translates to a sample of 37,960 observations out of the benchmark sample of 39,275 observation (96 percent).

I re-run the benchmark estimates of Table 2.3 using these population-weighted covariates and report the results in Table 2.5. These home language "country" estimates are very similar to the benchmark estimates, though the positive effect of genetic distance drops slightly in magnitude. The stability of the coefficient estimates indicates that the benchmark result is not a consequence of the *Ethnologue* home country assignment rule. The fact that the opposing forces of linguistic and genetic distance are borne out of these alternative assignment estimates is quite reassuring given that the set of transformed covariates all exhibit a significant difference in means relative to the comparable measures in the benchmark sample (Table C9). The similarity of these estimates to the benchmark estimates ensure the *Ethnologue* assignment rule is a good approximation of a book translation's country of origin.

2.4.2 Human Capital

A country's decision to translate might also be influenced by its level of human capital, since educated populations are more likely to have a high demand for book translations. Using data from Barro and Lee (2013), I measure human capital in the translating country in four ways: a country's average years of schooling attained, the percentage of the total population who have completed primary schooling, secondary schooling and tertiary schooling. Because these data are only available in 5-year intervals I supplement it with data from the Penn World Tables version 8.0. I use a yearly index of human capital per person, based on years of school (Barro and Lee, 2013) and the returns to education (Psacharopoulos, 1994). Table 2.6 reports robustness checks using these different measures of human capital.

Column (1) includes the Penn World Table human capital index, which I match to 99 percent of

[21] Linguistic distance is measured between the target and original language and thus measured independently of the home country assumption.

[22] The results are also robust to both stricter and weaker restrictions on coverage.

Table 2.6: Robustness Check: Human Capital and Education

	Dependent variable: Log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	-1.14***	-0.95***	-0.95***	-0.96***	-0.95***	-0.96***
	(0.29)	(0.31)	(0.31)	(0.31)	(0.31)	(0.31)
Genetic distance	2.38**	3.13***	3.12***	3.12***	3.10***	3.12***
	(1.05)	(1.18)	(1.18)	(1.18)	(1.18)	(1.18)
Geographic distance	-1.38***	-1.08**	-1.09**	-1.09**	-1.08**	-1.09**
	(0.46)	(0.46)	(0.46)	(0.47)	(0.47)	(0.47)
Human capital index translating country	0.13					
	(0.14)					
Average years of schooling translating country		0.06*				
		(0.03)				
% of primary schooling translating country			-0.01*			0.00
			(0.00)			(0.00)
% of secondary schooling translating country				0.01**		0.01*
				(0.00)		(0.00)
% of tertiary schooling translating country					0.01	0.02
					(0.01)	(0.01)
Benchmark controls	Yes	Yes	Yes	Yes	Yes	Yes
Translating Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	38,907	8,815	8,815	8,815	8,815	8,815
Adjusted R^2	0.28	0.26	0.26	0.26	0.26	0.26
Country pair clusters	1809	1342	1342	1342	1342	1342
Standardized Coefficients (%)						
Linguistic Distance	-8.0	-6.8	-6.8	-6.8	-6.7	-6.8
Genetic Distance	6.2	8.1	8.1	8.0	8.0	8.0

This table tests for selection on the human capital level of a translating country. All regressions include the benchmark set of control variables used in column (6) of Table 2.3. Country-pair clustered robust standard errors reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.7: Robustness Check: Unobserved Country-Pair Effects

	Dependent variable: Log translations			
	(1)	(2)	(3)	(4)
Linguistic distance	-0.86*	-0.89*	-0.88*	-0.88*
	(0.47)	(0.52)	(0.52)	(0.52)
Economic controls	No	No	Yes	Yes
Political controls	No	No	Yes	Yes
Trade controls	No	No	No	Yes
Translating Language FE	Yes	Yes	Yes	Yes
Country Pair FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	42,415	38,928	38,928	38,928
Adjusted R^2	0.36	0.35	0.35	0.35
Country pair clusters	1711	1551	1551	1551
Standardized Coefficients (%)				
Linguistic Distance	-6.3	-6.3	-6.3	-6.3

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

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

2.4.3 Existing Bilateral Relationships

The most effective way to capture any relevant time-invariant feature of a country pair is to include country pair fixed effects. The major limitation of this approach is that it is no longer possible to explicitly estimate genetic distance because country pair variation is time-invariant and adsorbed by the fixed effects estimator.[23] Nonetheless I proceed to test the robustness of the benchmark linguistic distance estimate.

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

Table 2.7 reports the benchmark estimate with country pair fixed effects in place of individual country effects. Column (1) reports estimates where I forego any covariates in order to maximize the sample. Linguistic distance is estimated to be significantly different from zero with the expected sign and with a magnitude of influence similar to the benchmark estimate. I then incrementally add the benchmark covariates that are not absorbed by the country-pair fixed effects. Overall the results are similar to the benchmark, albeit a little noiser and consequently less precisely estimated. However, the robustness of the linguistic distance estimate to country-pair fixed effects confirms that time-invariant country-pair variation cannot explain away the benchmark estimate.

2.4.4 Check for Understated Standard Errors

Because linguistic and genetic distance are assumed to be constant over the sampled time period, standard errors may be understated due to repeated values of each distance measure for the same language-country pair in the panel (Bertrand et al., 2004). To test for this I regress log translations on a set of year dummies and collapse the data by averaging over time the residual translation variation. Similarly, all other time-variant independent variables are regressed on time dummies and the residuals are collapsed by averaging over time. The benchmark specification is then re-estimated using this collapsed and time-averaged data. Results are reported in table 2.8.

Linguistic and genetic distance are again estimated significantly significant and have the expected signs. The only feature of these results that is different from the benchmark estimate is the standardized coefficient for genetic distance is greater in absolute magnitude than linguistic distance. However, the main finding of the opposing forces of linguistic and genetic distance remains unchanged.

2.4.5 Differences in Across Country Language Structure

One additional concern is that countries are structurally different in terms of language. For example, multilingual countries may translate more books because they are more linguistically diverse. Similarly, the colonial legacy of language is evident in the fact that many countries still use their colonizer's native language as a regional lingua franca. This may influence both the number of books translated and specific languages that are commonly translated. A shared language between two countries may also influence how much those two countries translate each other's writings. In this section, I investigate further sample restrictions to be sure the benchmark results are not driven by country-level differences in language structure. Results are reported in Table 2.9, where all estimates include a full set of controls.

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

Next I investigate the influence of lingua francas within a translating country. Using data

Table 2.8: Robustness Check for Understated Standard Errors

	Dependent variable: Time averaged log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance	-0.63***	-0.66***	-0.66***	-0.58***	-0.59***	-0.57***
	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)	(0.16)
Genetic distance	1.73***	1.79***	1.78***	1.83***	1.83***	1.69***
	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)	(0.47)
Geographic distance	-0.19***	-0.21***	-0.21***	-0.11	-0.12*	-0.35**
	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.17)
Economic controls	No	Yes	Yes	Yes	Yes	Yes
Political controls	No	No	Yes	Yes	Yes	Yes
Trade controls	No	No	No	Yes	Yes	Yes
Colonial controls	No	No	No	No	Yes	Yes
Geography controls	No	No	No	No	No	Yes
Target Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Target Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,084	7,084	7,084	7,084	7,084	7,084
Adjusted R^2	0.10	0.11	0.11	0.11	0.11	0.11
Country pair clusters	1891	1891	1891	1891	1891	1891
Standardized Coefficients (%)						
Linguistic Distance	-7.9	-8.3	-8.2	-7.3	-7.3	-7.0
Genetic Distance	9.2	9.5	9.4	9.7	9.7	8.9

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

Table 2.9: Robustness Check for Differences in Across-Country Language Structure

	Dependent variable: Log translations				
	(1)	(2)	(3)	(4)	(5)
Excluding:	Multilingual Countries	Lingua Franca Translations	Common Language Countries (official)	Common Language Countries ($> 9\%$ pop.)	All Excluded Countries
Linguistic distance	-1.18***	-1.10***	-1.43***	-1.41***	-1.37***
	(0.32)	(0.28)	(0.30)	(0.31)	(0.33)
Genetic distance	2.82**	2.43**	2.91***	2.92***	2.63**
	(1.13)	(1.07)	(1.07)	(1.07)	(1.13)
Geographic distance	-1.26***	-1.27***	-1.44***	-1.23***	-1.39***
	(0.49)	(0.45)	(0.42)	(0.42)	(0.46)
Benchmark controls	Yes	Yes	Yes	Yes	Yes
Translating Language FE	Yes	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	32,588	37,628	34,481	33,589	28,279
Adjusted R^2	0.29	0.29	0.31	0.31	0.31
Country pair clusters	1636	1840	1743	1707	1440
Standardized Coefficients (%)					
Linguistic Distance	-8.2	-7.8	-10.0	-9.8	-9.4
Genetic Distance	7.3	6.3	7.3	7.4	6.5

This table reports estimates from various subsamples that exclude potentially problematic countries and translations. All regressions include the benchmark set of control variables used in column (6) of Table 2.3. Country-pair clustered robust standard errors reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

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

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

2.5 Distance Effects by Idea Types

This section extends the analysis by investigating if the two opposing forces of relatedness exist across different idea types. I narrow the focus to two categories of idea type: economic ideas and cultural ideas. I define a book as having economic use value when it is scientific in nature, including translations pertaining to the applied sciences and natural sciences. Conversely I define a book to have cultural use value when a translation pertains to the field of social science, philosophy, history, literature, religion or the arts.[24]

To preserve the sample size of the benchmark estimates I construct the dependent variable for economic book translations as $\ln(1 + \text{economic translations})$ and the dependent variable for cultural book translations as $\ln(1 + \text{cultural translations})$. In other words, in instances where no book translations of an economic idea type are observed, the dependent variable for economic ideas equals zero where it would have otherwise taken a strictly positive value in the benchmark estimates. The opposite is true for the measure of cultural book translations.

Table 2.10 reports estimates for each idea type. Consistent with the benchmark estimates, both linguistic and genetic distance are significantly estimated with the expected sign irrespective of idea type. These results suggest the opposing forces of relatedness are still at play when disaggregating books by their economic and cultural use value.

More interesting is the fact that economic book translations are more responsive to genetic differences than cultural book translations, as indicated by the standardized betas. Ashraf and Galor (2013) establish that the positive effect of population diversity on income per capita results from a wider spectrum of traits found amongst a diverse group, which increases the likelihood

[24] I also separately drop observations for each subject of translation to test the sensitivity of the benchmark estimate to different idea types. The benchmark results are unaltered (see Appendix Table C8).

Table 2.10: Distance Effects by Cultural and Economic Idea Types

	(1)	(2)	(3)	(4)
	Cultural Book Translations		Economic Book Translations	
Linguistic distance	-1.19***	-0.86***	-1.04***	-0.77***
	(0.24)	(0.24)	(0.22)	(0.21)
Genetic distance	2.69***	1.93**	2.22***	1.67**
	(0.88)	(0.88)	(0.77)	(0.73)
Geographic distance	-0.73***	-1.28***	-0.60***	-0.68**
	(0.13)	(0.38)	(0.08)	(0.31)
Benchmark controls	No	Yes	No	Yes
Translating Language FE	Yes	Yes	Yes	Yes
Translating Country FE	Yes	Yes	Yes	Yes
Original Country FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	39,275	39,275	39,275	39,275
Adjusted R^2	0.26	0.27	0.25	0.26
Country pair clusters	1897	1897	1897	1897
Standardized Coefficients (%)				
Linguistic Distance	-10.0	-7.3	-11.8	-8.7
Genetic Distance	8.3	6.0	9.2	7.0

This table establishes that the two opposing forces of relatedness exist across different idea types. Cultural book translations are those from the field of philosophy, social sciences, history, literature, religion and the arts. Economic book translations are those from the fields of natural and applied sciences. All regressions include the benchmark set of control variables used in column (6) of Table 2.3. Country-pair clustered robust standard errors reported in parentheses. * p $<$ 0.10, ** p $<$ 0.05, *** p $<$ 0.01.

of complementary traits and thus aggregate productivity. In the context of this paper, genetically distant groups likely possess a wider spectrum of traits than similar groups, which increases the likelihood of complementary ideas and therefore an incentive for communication exists when groups have more to learn from each other. Hence the incentive to translate economically relevant books is greater than culturally relevant books because the diversity-driven incentive for idea flows is linked to aggregate productivity.

2.6 Concluding Remarks

This paper is motivated by the observation that linguistic and genetic distance can account for contemporary cross-country income differences. One interpretation of this evidence is that these distance measures proxy as indices of population relatedness, and that cross-country communication and adoption of ideas is more likely when two countries are closely related (Spolaore and Wacziarg, 2009, 2013a). To test this interpretation I use data on book translations as a measure of international idea flows. I empirically establish that idea flows exhibit a trade-off between two

opposing forces of population relatedness: linguistic differences impose a cost on idea flows while genetic differences provide a communication incentive. This speaks to the evidence of two opposing forces of relatedness found to exist *within* countries (Ashraf and Galor, 2013), and contributes to this literature with evidence that these opposing forces of relatedness also exist *between* countries.

To reconcile this evidence with the interpretation of Spolaore and Wacziarg (2009) I also show that linguistic distance reflects a stronger relationship with book translations (in absolute terms) than genetic distance. Similarly geographic distance reflects a large and robust negative relationship with book translations. This suggests that when treating genetic distance as a summary measure of population differences, it should reflect a negative relationship with book translations because of the latent linguistic and geographic variation captured by unconditional genetic distance. This is indeed what I find, and that only after accounting for linguistic and geographic distance do I observe a stable positive relationship between genetic distance and book translations.

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

Chapter 3

Ecology, Trade and the Geographic Origins of Ethnolinguistic Differences

3.1 Introduction

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; Ahlerup and Olsson, 2012). Yet this cultural affinity can also form a rift between dissimilar ethnic groups.

Ashraf and Galor (2013) provide evidence of a trade-off between the beneficial and detrimental effects of ethnic diversity at the national level, and corroborate this evidence at the subnational level (Ashraf et al., 2015). Ethnic diversity also explains cross-country differences in public policy and income (Easterly and Levine, 1997; Alesina et al., 2003), institutions (La Porta et al., 1999), the prevalence of conflict (Esteban et al., 2012) and more (see Alesina and La Ferrara (2005)). Related to this is the effect of ethnic or genetic distance between groups on cross-country income (Spolaore and Wacziarg, 2009), conflict (Spolaore and Wacziarg, 2016), in addition to within-country patterns of ethnic favoritism (Dickens, 2016a), inequalities in child mortality rates (Gomes, 2014) and more (see Spolaore and Wacziarg (2013a)). If the historical determinants of ethnic group differences are not well understood, then their lingering effects in the present may go unnoticed or be incorrectly attributed to other channels of influence.

In this paper, I empirically examine the geographic origins of ethnic differences, and offer an economic interpretation as to why we observe these differences in similarity. This research is motivated by a growing literature that aims to correctly identify the historical determinants of group differences, and to understand the specific transmission mechanisms that account for this historical persistence (Nunn, 2013; Spolaore and Wacziarg, 2013a). It has been shown that cultural values and human capital skills are transmitted through migration (Putterman and Weil, 2010; Easterly and Levine, 2012), colonialism (Nunn, 2008; Nunn and Wantchekon, 2011), institutions

(Alesina and Giuliano, 2013) and more, but there is a scarcity of evidence that explains why ethnic diversity and ethnic differences exist in the first place.

Michalopoulos (2012) shows that regions characterized by homogeneous land endowments tend to be more homogeneous in ethnicity, and consequently less diverse today. The underlying mechanism is that homogeneous land endowments historically gave rise to the same set of location-specific skills among inhabitants of that region, resulting in the emergence of localized ethnicities. Ahlerup and Olsson (2012) also study the origins of ethnic diversity, and similarly find that isolating geographic features tend to increase ethnic diversity, but they also show that ethnic diversity is particularly pronounced in countries where humans settled relatively early.[1]

Yet this scarce but growing body of evidence on the origins of ethnic diversity only speaks to the why some regions are more diverse than others, and cannot speak to why some ethnic groups are relatively more different than others. Hence, a subtle yet important question remains unanswered: why are some ethnic groups more dissimilar from each other than others? This research constitutes the first attempt to fill this gap by shedding light on the geographic origins of ethnolinguistic distance.

I propose that trade is one factor driving this variation in ethnic distance between neighbouring groups. The incentive for farmers to specialize in crop production is larger in agriculturally diverse regions, where trade results as a necessary outcome of specialization. This proposition speaks to the work of Bates (1983), who posits a theory of state formation in agriculturally diverse regions. The crux of his argument is that trade was more profitable – and thus more frequent – *across* ecological boundaries because of the diversity of tradable agricultural goods.[2] Because groups would historically trade by interacting and communicating with each other, the extent of drift is mitigated by frequent interaction. Hence, the resulting inter-group trade is the channel through which agricultural diversity mitigates cultural drift.

Implicit in Bates' theory, is this idea that the agricultural diversity of a particular region is a proxy for the historical gains from trade in that region. In a related paper, Fenske (2014) uses agricultural diversity to approximate the historical gains from trade. Indeed, Bates (2010, p.21) reiterates this point when discussing the historical development of agrarian societies: "The diversity of the ecosystem thus promotes diversity in production and, with it, exchange over space."

To examine this empirically I study the geo-climatic conditions that separate neighbouring ethnolinguistic groups. To do this I use GIS software to identify all neighbouring ethnolinguistic groups around the world, and create a 100 km buffer zone around the border segment connecting a neighbouring ethnolinguistic pair (i.e., a 50 km zone that follows the border in each group). I measure the agricultural diversity of a buffer zone as the variation in potential agricultural output. Because these border zones delineate ethnolinguistic pairs that vary in their linguistic distance, I can empirically test the effect of agricultural diversity on cultural drift as reflected by the similarity

[1] In a recent manuscript, Galor et al. (2016) find that regional variations in geographic characteristics explain cross-language variation in linguistic structures, in particular the presence of the future tense.

[2] In particular, Bates (1983) argues that state formation was a natural response to the uncertainty of trade in a world without protected property rights. Fenske (2014) provides empirical evidence of this theory.

of language between neighbouring groups.

The second part of my empirical strategy accounts for the fact that neighbouring ethnolinguistic groups do not always descend from the same language family. In such instances the distance between two languages might simply reflect a different ancestry. As a solution to this problem I use the structure of language trees and narrow my focus to "sibling" ethnolinguistic pairs – those that descend from the same parent language. I define these pairs to be siblings because they share an identical ancestral history, separated only at the most recent cleavage on the *Ethnologue* language tree. Narrowing the focus to sibling pairs implicitly accounts for all unobserved historical characteristics of a language family that might plausibly affect the drift between two groups. Doing so thus disentangles the effects of shared ancestry from the effect of geo-climatic characteristics that I'm interested in.

The third part of my empirical strategy exploits the exogenous change in potential agricultural diversity that was the result of 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 source of historical variation in a region's potential output and thus agricultural diversity (Galor and Özak, 2016).

Consistent with my hypothesis, results indicate that greater agricultural diversity reduces the linguistic distance between neighbouring groups. This evidence is consistent across the full sample and the sibling sample. These results are robust to the inclusion of numerous geographic, climatic and disease environment control variables.

In addition to my main finding, I show that ruggedness and the prevalence of malaria in border regions amplify the distance between groups, whereas a lake or river within the border region brings groups closer together. I also find that groups neighbouring across a north-south axis orientation tend to be more ethnically distant, which is consistent with Blouin (2014), who empirically verifies Diamond's (1997) hypothesis that latitudinal differences reduced the historical flow of information and thus communication between groups.

This research provides the first attempt to document the geographic origins of ethnolinguistic distance. My main finding and contribution is the empirical evidence that variations in pre-industrial agricultural diversity are at the root of existing cross-group variations in linguistic differences. Although I do not have georeferenced data on the historical gains from train, the proposed trade mechanism is consistent with the historical and anthropological evidence from Bates (1983). The corollary contribution is thus an empirical verification of Bates's (1983) theory.

The rest of this paper is organized as follows. Section 3.2 summarizes the Columbian Exchange and the expansion of available agricultural goods. Section 3.3 presents the data and how the buffer zones are constructed. Section 3.4 outlines the empirical model and identification strategy, and presents the empirical results. Section 3.5 concludes.

3.2 The Columbian Exchange

The Columbian Exchange refers to the time period following the encounter of the Americas in 1492 by Christopher Columbus. In particular the widespread exchange of crops, disease, ideas, populations and more between the Old and New World. Nunn and Qian (2010) summarize the existing literature on the Columbian Exchange and provide insight into less-studied areas of the exchange.

This coming together of the continents introduced a new set of available crops for cultivation that, in the context of this paper, provide an exogenous source of variation in potential agricultural diversity. Both the availability of new crops and the availability of unpopulated land suitable for agriculture in the New World had a substantial impact on the global supply of agricultural goods (Nunn and Qian, 2010).

The natural experiment associated with the Columbian Exchange is the unexpected change in a tract of land's potential output following the expansion of available agricultural goods. Galor and Özak (2016) use this source of variation to study the effect of agricultural returns on contemporary time preference at the individual, regional and national level. Nunn and Qian (2011) also use time variation coming from the Columbian Exchange with regional variation in suitability for potato cultivation to estimate the effect of the potatoes introduction to the Old World on population and urbanization rates.

3.3 Data

3.3.1 Linguistic Distance

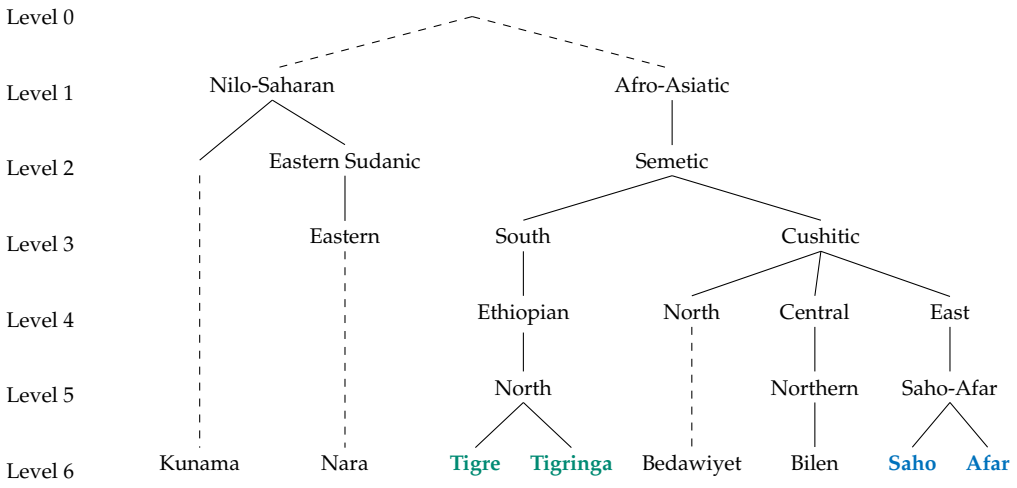
As my outcome variable I use a computerized lexicostatistical measure of linguistic distance developed by the Automatic Similarity Judgement Program. This lexicostatistical measure captures the phonological dissimilarity between two languages as an approximation of ethnic distance. See Dickens (2016a) for a discussion of this measure and Appendix A for a formal discussion of the estimation procedure.

Lexicostatistical distance is indispensable to this paper. Economists often use the ratio of shared branches on a language tree as a measure of proximity between languages – known as a cladistic distance.[3] However, when analysing sibling pairs there is no variation in cladistic distance because the two languages separated at the most recent cleavage on the *Ethnologue* language tree. In other words, each sibling pair share an identical number of branches on a language tree. To the contrary, lexicostatistical distance explicitly measures dissimilarity between two languages so there is no reason two separate sibling pairs should be identical in their distance.

Figure 3.1 makes this point clear. I've drawn the *Ethnologue* language tree for the 8 major Eritrea languages, and coloured coded sibling ethnolinguistic pairs. Each sibling pair share 5

[3] See Ginsburgh and Weber (2016) for a review of this literature.

Figure 3.1: Phylogenetic Tree of Eritrean Languages



This figure depicts the language tree for the 8 major languages of Eritrea. Among these 8 languages there are 2 language pairs that share a parent language on the language tree. I've colour coded these sibling ethnolinguistic pairs: Tigre-Tigringa in green and Saho-Afar in blue.

out of the 6 branches – the maximum number of shared branches for two distinct languages.[4] Because the ratio is always identical across sibling pairs, there is no observable variation among these pairs. To the contrary, the lexicostatistical estimate for the Tigre-Tigringa sibling pair is 63.2 percent dissimilarity whereas the estimate for the Saho-Afar sibling pair is only 50.9 percent dissimilarity.

3.3.2 Independent Variables

Unit of Observation: Buffer Zones

I use the *Ethnologue's* (16^{th} edition) mapping of ethnolinguistic groups to construct 100 kilometre buffer zones that follow the length of a border segment delineating each neighbouring ethnolinguistic pair.[5] By construction these buffer zones provide a lens to examine the geography of border regions, extending 50 kilometres in every direction from each point on the border segment delineating a neighbouring pair. Throughout the empirical analysis, these buffer zones serve as my unit of observation.[6]

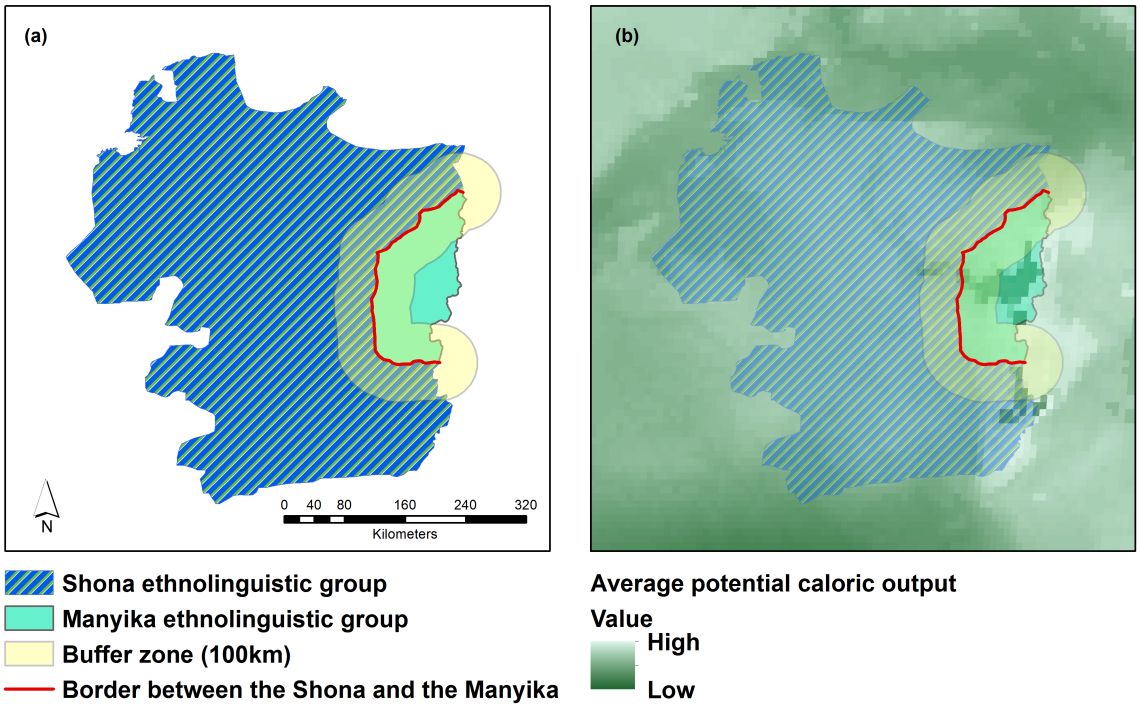
Consider, as an example, the neighbouring Shona and Manyika groups in Zimbabwe. Panel (a) of Figure 3.2 depicts the spatial distribution of these ethnolinguistic groups, and the red line

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

[5]I use the World Goode Homolosine projection because it is an equal-area projection that minimizes area distortion in measurement.

[6]I impose two size restrictions on a valid sample observation: each ethnolinguistic group's total area must be at least 100 square kilometres in size and must have a population of speakers greater than 100 people. However, both restrictions are innocuous and do not change the interpretation of the estimates in any way.

Figure 3.2: Example: Buffer Zone Unit of Observation



represents the segment of border delineating the Shone and Manyika. I then construct a Euclidean buffer zone that extends 50 kilometres from every point on the border, resulting in a total buffer zone width of 100 kilometres. In effect, the buffer zone is a perimeter around the delineating segment of border that I use to measure different geographic characteristics contained within the enclosed region. Panel (b) exemplifies how I overlay the Shona-Manyika buffer zone onto raster data and measure different geographic characteristics within the buffer region.

Caloric Suitability Index and Agricultural Diversity

As a measure of agricultural diversity I use variation in the Caloric Suitability Index (CSI) of each buffer zone. Galor and Özak (2016) introduce this novel measure of agricultural productivity – a standardized measure of productivity based on the agro-climatic determinants of the *potential* caloric output of each $5' \times 5'$ grid cell on earth.[7] Data for this measure comes from the Global Agro-Ecological Zones (GAEZ) project of the Food and Agriculture Organization (FAO).

Galor and Özak (2016) use the FAO's estimate of potential crop yield for the 48 available crops under low-level inputs and rain-fed agriculture to construct a composite measure of potential output in each grid cell. To account for differences in the nutritional value of different crops, they convert the estimated potential output for each crop to reflect its potential caloric return. They then average across crops within each $5' \times 5'$ grid cell to construct an average measure of caloric suitability.

[7] http://ozak.github.io/Caloric-Suitability-Index/

When constructing this average measure, Galor and Özak (2016) make an important distinction between the caloric potential of a grid cell in the pre-1500 and post-1500 period. The difference between these two different measures reflects the change in potential caloric output that resulted from the expansion of agricultural goods during the Columbian Exchange.

To measure agricultural diversity, I overlay a map of the constructed buffer zones onto the CSI raster data, and calculate the standard deviation of both the pre-1500 and post-1500 CSI within each buffer zone. I also calculate the average of the CSI within each buffer zone.

As an alternative measure of agricultural diversity, I use Ramankutty et al.'s (2002) Agricultural Suitability Index (ASI) data. Similar to my calculations with the CSI data, I use the standard deviation of the ASI within a border region as a measure of agricultural diversity. However, the ASI is not available for the pre-1500 period, so it is not possible to exploit the expansion of agricultural goods during the Columbian Exchange. See Galor and Özak (2015) for a more detailed discussion of the additional advantages of the Caloric Suitability Index over Ramankutty et al.'s (2002) Agricultural Suitability Index.

Additional Geo-Climatic and Disease Environment Data

In addition to the data on agricultural productivity, I collect a variety of other geographic and climatic data. Data on average temperature and precipitation come from the WorldClim – Global Climate Database. These datasets provide information on temperature and precipitation at a spatial resolution of $5' \times 5'$. I use these raster data in an identical way to the CSI data; I calculate both the mean and standard deviation of the data within each buffer zone.

I use elevation data from the National Oceanic and Atmospheric Adminstration (NOAA). These data are available at a 30-arc-second resolution. In addition to measuring the level of elevation within each buffer zone, I calculate the standard deviation of elevation as a measure of ruggedness (Michalopoulos, 2012; Kitamura and Lagerlof, 2016).

I also construct dummy variables for the presence of a lake and river in the buffer zone. These data come from georeferenced maps from Natural Earth, which I intersect with my map of buffer zones to indicate border regions with either a lake or river.

As a measure of the disease environment, I use the Malaria Ecology Index to construct a measure of the average prevalence of malaria in a buffer zone. These data come from Kiszewski et al. (2004) and are available at a spatial resolution of $0.5° \times 0.5°$.[8]

3.3.3 Does Geography Delineate Ethnolinguistic Groups?

Before proceeding to the empirical analysis, I check if there is something unique about the geography of ethnolinguistic group border regions. Michalopoulos' (2012) principal finding is that ethnolinguistic groups form across homogeneous geographic terrain. By extension this finding

[8]See Table D1 and Table D2 in the appendix for the full sample and sibling sample summary statistics.

suggestions that group boundaries should exhibit terrain distinct from the interior of the associated group. To test for this possibility I look at differences in means for a number of geographic, climatic and disease environment variables. Table 3.1 reports these mean differences. I find that for in all but 2 instances there is a statistically significant difference between the interior and the border region of each group. This suggests there is something unique about the geography of the border regions that separate groups, an observation consistent with Michalopoulos (2012).

3.4 Empirical Strategy and Estimates

3.4.1 Identification Strategy

To overcome possible threats to identification I take a number of steps. First, in all regressions I include continental and language family fixed effects. Doing so not only accounts for unobserved confounding factors at the continental level, but also time-invariant family-specific unobservables that may have influenced the historical sorting and interaction of ethnolinguistic groups.

Second, I take into account the fact that neighbouring ethnolinguistic groups do not always descend from the same language family by narrowing my focus to sibling pairs. By construction the sibling analysis imposes the restriction that a neighbouring pair must descend from the same parent language and thus implicitly accounts for the entire ancestral history of a sibling pair. By doing so I disentangle the effect of shared ancestry from the effect of agricultural diversity on the linguistic distance between neighbouring groups.

Third, I exploit variation in potential agricultural diversity using changes in the available set of crops for cultivation resulting from the onset of the Columbian Exchange. A threat to identification is the possibility that ethnolinguistic groups defined group borders along agriculturally diverse regions in the pre-1500 period because of the capacity of these regions to sustain large populations. Should this be the case then the location of group borders would be the result of endogenous sorting of groups. Identification of agricultural diversity then rests on the assumption that the change in the set of available crops resulting from the Columbian Exchange is random and independent of all other factors in a buffer zone, conditional on the set of pre-1500 cultivatable crops.

3.4.2 Empirical Model and Results

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

$$LD_k = \beta_0 + \beta_1^{1500} AD_k + \beta_1^{ch} \Delta AD_k + x_k' \Phi + \gamma_l + \delta_c + \epsilon_k. \tag{3.1}$$

The dependent variable LD_k measures the linguistic distance between neighbouring ethnolinguistic groups i and j in buffer zone k. AD_k denotes pre-1500 CSI variation (agricultural diversity)

Table 3.1: Difference in Means: Language Pair Zone vs. Language Pair Buffer Zone

	Full Sample			Sibling Sample		
	Border Buffer Zone	Lang. Pair Zone	Standard Error	Border Buffer Zone	Lang. Pair Zone	Standard Error
CSI (pre-1500, Avg.)	1.346	1.360	(0.004)***	1.383	1.381	(0.007)
Change in CSI (post-1500, Avg.)	-0.061	-0.063	(0.002)	-0.119	-0.106	(0.004)***
CSI Variation (pre-1500, Avg.)	0.222	0.217	(0.002)**	0.266	0.185	(0.006)***
Change in CSI Variation (post-1500, Avg.)	-0.063	-0.045	(0.002)***	-0.110	-0.057	(0.004)***
Elevation	0.696	0.712	(0.003)***	0.716	0.744	(0.006)***
Ruggedness	0.299	0.263	(0.002)***	0.337	0.249	(0.005)***
Malaria	8.038	8.156	(0.039)***	8.800	9.034	(0.108)**
Precipitation	13.65	13.52	(0.021)***	15.24	15.32	(0.041)*
Precipitation Variation	1.683	1.490	(0.015)***	1.879	1.225	(0.033)***
Temperature	21.54	21.73	(0.031)***	22.17	22.29	(0.066)*
Temperature Variation	1.552	1.414	(0.012)***	1.734	1.272	(0.026)***

This table establishes that the mean value of these geo-climatic factors measured along the border regions delineating contiguous language pairs are statistically distinct from the mean value of these factors throughout the total language pair area. In other words, the statistically significant difference in means suggests there is something unique of about geo-climatic features of border regions delineating contiguous language pairs. The full sample difference in means have a sample size of $n = 6,990$ while the sibling sample refers to a sample size of $n = 1,277$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

in buffer zone k, and ΔAD_k denotes the change in CSI variation (change in agricultural diversity) following the onset of the Columbian Exchange. x_k represents a vector of buffer zone geo-climatic characteristics,[9] γ_l is a complete set of language family fixed effects for groups i and j, and γ_c is a complete set of continental fixed effects associated with the buffer zone region k.

Unconditional Estimates

Table 3.2 presents the unconditional estimates of equation (3.1). Column 1 reports estimates for the average value of pre-1500 CSI and the change in average CSI. Neither estimate is statistically significant. However these measures reflect the average potential productivity on a region and not the diversity.

Column 2 reports estimates for the pre-1500 variation in CSI and the change in that variation at the onset of the Columbian Exchange. These measures capture the agricultural diversity of a buffer zone. Both estimates enter with the expected negative sign and are statistically significant at the 1 percent level. This says that neighbouring groups are more similar in their language when the region delineating those groups exhibits greater diversity in potential agricultural output.

Column 3 includes both the potential productivity measures of column 1 and the potential diversity measures of column 2. Reassuringly, the estimates for pre-1500 CSI variation and the change in CSI variation remain remarkably stable in magnitude and significance even when accounting for the average potential productivity of a buffer zone. The estimates in column 4 include continent fixed effects, and although the magnitude of the estimates drop, the effect remains statistically significant at the 10 percent level.

Columns 5-8 report the analogue estimates of column 1-4 using a sample of sibling pairs. The estimates in column 5 are almost twice that of the full sample estimates in column 1. For the change the CSI the effect is now estimated to be significant at the 5 percent level. The estimates in column 6 are similarly much larger in magnitude with little change in the estimated standard error. However, as column 7 makes clear, the significant effect of the change in CSI goes away once CSI variation and the change in CSI variation are included the estimating equation.

All together the estimates of Table 3.2 suggestion two things: that ethnic groups separated across agriculturally diverse regions are more similar in language than groups separated across homogeneous regions, and the precision of the estimates greatly improve after accounting for the ancestral history of an ethnolinguistic pair with the sibling-level analysis.

Conditional Estimates

Table 3.3 reports the conditional estimates of equation (3.1) on the sibling-level sample. The estimates are sorted in pairs, where an even-numbered column is identical to the previous odd-

[9]Including the pre-1500 potential crop yield and the change in that yield following the Columbian Exchange, the malaria suitability index, elevation, ruggedness, precipitation and precipitation variation, temperature and temperature variation, a dummy variable for the presence of a lake, a dummy variable for the presence of a river, the total area of an ethnolinguistic pair, the latitudinal difference between group centroids, the longitudinal difference between group centroids, the interaction of the latitude and longitudinal differences, and variation in land suitability for agriculture.

Table 3.2: Border-Level Regressions: Unconditional Caloric Suitability Index Benchmark Results

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$								
	Full Sample				Sibling Sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CSI (pre-1500)	-0.007		-0.005	-0.007	-0.017		-0.017	-0.026*
	(0.007)		(0.007)	(0.008)	(0.014)		(0.014)	(0.014)
Change in CSI (post-1500)	-0.019		-0.010	0.009	-0.039**		-0.007	0.004
	(0.013)		(0.014)	(0.016)	(0.017)		(0.018)	(0.022)
CSI Variation (pre-1500)		-0.067***	-0.065***	-0.059**		-0.105***	-0.103***	-0.117***
		(0.023)	(0.023)	(0.024)		(0.036)	(0.036)	(0.037)
Change in CSI Variation (post-1500)		-0.113***	-0.105***	-0.076*		-0.218***	-0.223***	-0.197***
		(0.036)	(0.039)	(0.041)		(0.049)	(0.052)	(0.055)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Continental FE	No	No	No	Yes	No	No	No	Yes
Language Cluster 1	2,713	2,713	2,713	2,713	942	942	942	942
Language Cluster 2	1,769	1,769	1,769	1,769	851	851	851	851
Observations	6,990	6,990	6,990	6,990	1,277	1,277	1,277	1,277
Adjusted R^2	0.26	0.26	0.26	0.26	0.32	0.34	0.34	0.35

This table establishes the negative and statistically significant effect of the variation in caloric suitability on a language pair's lexicostatistical linguistic distance. The unit of observation is a 100 km buffer zone along the contiguous border segment of each language pair. 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 3.3: Border-Level Regressions: Conditional Caloric Suitability Index Benchmark Results

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CSI Variation (pre-1500)	-0.103***	-0.117***	-0.167***	-0.185***	-0.169***	-0.195***	-0.173***	-0.199***
	(0.036)	(0.037)	(0.064)	(0.063)	(0.065)	(0.064)	(0.064)	(0.063)
Change in CSI Variation (post-1500)	-0.223***	-0.197***	-0.254***	-0.238***	-0.237***	-0.233***	-0.255***	-0.257***
	(0.052)	(0.055)	(0.066)	(0.068)	(0.066)	(0.069)	(0.065)	(0.068)
CSI (pre-1500)	-0.017	-0.026*	-0.016	-0.029*	-0.012	-0.026*	-0.017	-0.029*
	(0.014)	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)	(0.015)	(0.016)
Change in CSI (post-1500)	-0.007	0.004	-0.011	-0.006	-0.006	-0.007	-0.005	-0.008
	(0.018)	(0.022)	(0.019)	(0.024)	(0.019)	(0.025)	(0.019)	(0.025)
Elevation			0.021	0.022	0.023	0.021	0.021	0.020
			(0.019)	(0.019)	(0.018)	(0.018)	(0.018)	(0.018)
Ruggedness			0.198**	0.190**	0.193**	0.190**	0.175*	0.173*
			(0.096)	(0.094)	(0.096)	(0.095)	(0.096)	(0.096)
Precipitation			-0.000	-0.000	-0.001	-0.001	-0.001	-0.001
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Precipitation Variation			0.002	0.001	0.003	0.003	0.002	0.002
			(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Temperature			0.002	0.003*	0.002*	0.003**	0.003*	0.003**
			(0.002)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
Temperature Variation			-0.029	-0.026	-0.029	-0.026	-0.026	-0.024
			(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
River			-0.037***	-0.040***	-0.031***	-0.034***	-0.032***	-0.035***
			(0.012)	(0.012)	(0.011)	(0.011)	(0.011)	(0.011)
Lake			-0.051**	-0.051**	-0.051**	-0.049**	-0.050**	-0.048**
			(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Malaria			0.003***	0.002*	0.002**	0.001	0.002*	0.001
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Area of Language Pair (km^2)					-0.009	-0.008	-0.010	-0.008
					(0.006)	(0.006)	(0.006)	(0.006)
Population of Language Pair					-0.004	-0.005	-0.004	-0.005
					(0.004)	(0.004)	(0.004)	(0.004)
Latitude Difference					0.011**	0.007	0.011**	0.007
					(0.005)	(0.006)	(0.005)	(0.006)
Longitude Difference					0.006	0.005	0.006	0.005
					(0.004)	(0.004)	(0.004)	(0.004)
Latitude $\times$ Longitude					-0.000	-0.000	-0.000	-0.000
					(0.001)	(0.000)	(0.001)	(0.001)
Land Suitability Variation							0.157**	0.138*
							(0.074)	(0.074)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Continental FE	No	Yes	No	Yes	No	Yes	No	Yes
Language Cluster 1	942	942	942	942	942	942	933	933
Language Cluster 2	851	851	851	851	851	851	841	841
Observations	1,277	1,277	1,277	1,277	1,277	1,277	1,267	1,267
Adjusted R^2	0.34	0.35	0.37	0.38	0.38	0.39	0.39	0.39

This table establishes the negative and statistically significant effect of the variation in caloric suitability on a language pair's lexicostatistical linguistic distance. The unit of observation is a 100 km buffer zone along the contiguous border segment of each language pair. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

numbered column but includes continental fixed effects. Overall it is clear that the estimates for CSI variation and change in CSI variation are quite robust to the inclusion of additional geo-climatic covariates, with or without continental fixed effects. In fact, adding these covariates increases the magnitude of the estimates. The estimates in column 8 imply that a one standard deviation increase in pre-1500 CSI variation reduced linguistic dissimilarity by 5.9 percentage points, while a one standard deviation increase in CSI variation in the post-1500 period reduced dissimilarity by an additional 6.1 percentage points.

A number of other covariates yield interesting results. I find that rugged border regions tend to amplify linguistic differences between sibling pairs. When a region possess substantial variation in elevation, communication is more costly and contact between neighbouring groups is limited. Thus, groups separated across rugged terrain tend to be more dissimilar in language than groups separated across homogeneous terrain. This intensive margin result is complimentary to the extensive margin evidence of Michalopoulos (2012), who finds that the number of ethnolinguistic groups is decreasing in a region's terrain ruggedness.

Another interesting result is the finding that the presence of a lake or river in a border region systematically limits the divergence of language between neighbouring groups. Lakes and rivers were historically focal points of commercial activity, and the resulting trade in these regions is a possible mechanism through which linguistic similarity is preserved.

To the contrary, I find greater dissimilarities between groups neighbouring across a north-south axis orientation. This is consistent with the evidence that latitudinal differences between groups reduced the historical flow of information and thus communication between groups (Diamond, 1997; Blouin, 2014). I also find that the high prevalence of malaria in a border region increases the linguistic distance between neighbouring groups. However, both these results are somewhat sensitive to the empirical specification, tending to fall short of statistical significant at conventional levels when including continental fixed effects.

Recent Migrations

In this paper, I propose a hypothesis that similar groups border agriculturally diverse areas because they have always lived there and their similarity is the result of inter-group trade. However, since the onset of the Columbian Exchange there has been considerable migration of populations across the world that might have led to the arrival of these groups in their current location because of a region's agricultural diversity. It's feasible that interconnected groups migrate to neighbouring regions, in particular agriculturally diverse regions that have the capacity to support a large number of people. Should this be true then it's not the case that agricultural diversity has any effect on the cultural similarity of the neighbouring groups.

To check for this I limit my sibling sample to pairs that reside in a country where a significant proportion of its population is native to that country. Data on the ancestral composition of a country's contemporary population comes from Putterman and Weil (2010). Table 3.4 reports estimates for a variety of specifications, where at least 25, 50 or 75 of the population is native to the

Table 3.4: Border-Level Regressions: Native Population Sensitivity Analysis

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$						
	Sibling Sample & >25% Native Pop.		Sibling Sample & >50% Native Pop.		Sibling Sample & >75% Native Pop.	
	(1)	(2)	(3)	(4)	(5)	(6)
CSI Variation (pre-1500)	-0.121***	-0.200***	-0.121***	-0.212***	-0.132***	-0.252***
	(0.037)	(0.065)	(0.037)	(0.065)	(0.039)	(0.070)
Change in CSI Variation (post-1500)	-0.209***	-0.248***	-0.218***	-0.269***	-0.202***	-0.281***
	(0.055)	(0.070)	(0.055)	(0.070)	(0.057)	(0.074)
Controls	No	Yes	No	Yes	No	Yes
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Continental FE	Yes	Yes	Yes	Yes	Yes	Yes
Language Cluster 1	915	915	886	886	767	767
Language Cluster 2	826	826	795	795	693	693
Observations	1,241	1,241	1,202	1,202	1,036	1,036
Adjusted R^2	0.34	0.39	0.34	0.39	0.33	0.37

This table tests the sensitivity of the benchmark estimates by limiting the sibling sample to pairs that reside in a country where a significant portion of their population is native to that country. Columns (1)-(2), (3)-(4) and (5)-(6) report estimates where at least 25, 50 and 75 percent of the population is native to the country of residence, respectively. Control variables are identical to those in columns (5) and (6) in Table 3.3. The unit of observation is a 100 km buffer zone along the contiguous border segment of each language pair. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

country of residence. In all cases the variables of interest are remarkably consistent in magnitude and significance. These estimates relieve any concern that the estimates are a spurious outcome of group migration.[10]

Overlapping and Disjoint Ethnolinguistic Groups

The *Ethnologue* mapping of ethnolinguistic groups poses two problems for this empirical exercise: group territories sometimes overlap and group territories are sometimes spatially disjoint. In the case of overlapping groups, this is problematic because a buffer zone will not be uniquely representative of the neighbouring pair. When a group is spatially disjoint (e.g., a single ethnolinguistic group occupies two non-adjacent regions within a country), each portion of the disjoint group will not necessarily be adjacent to the same set of neighbouring groups.

To address these concerns, I drop overlapping groups and groups composed of (disjoint) multi-parts from the sibling sample and re-estimate equation (3.1). Table 3.5 reports these estimates. Whether I drop overlapping groups (columns 1 and 2), multi-part groups (columns 3 and 4) or both (columns 5 and 6), the coefficients of interest remain statistically significant. Overall, the estimates in Table 3.5 imply that the main result of this paper is unaffected by the *Ethnologue's* mapping of groups.[11]

3.5 Concluding Remarks

This study takes the economic importance of ethnolinguistic distance as given, and goes a step deeper to explore the geographic foundation of these differences. I construct a novel georeferenced dataset to examine the border regions of neighbouring ethnolinguistic groups, together with variation in the set of potentially cultivatable crops at the onset of the Columbian Exchange, to estimate how agricultural diversity impacts linguistic differences between neighbouring groups. I find that ethnic groups separated across agriculturally diverse regions are more similar in language than groups separated across homogeneous agricultural regions.

I offer an explanation that speaks to the work of Bates (1983, 2010), who argues that trade was historically more profitable and thus more frequent in agriculturally diverse regions. Because groups would trade by interacting and communicating with each other, the extent to which their two languages drift apart is mitigated by frequent interaction. Hence, the resulting inter-group trade is the channel through which agricultural diversity mitigates ethnolinguistic differences.

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. It also suggests that the exogeneity of ethnolinguistic distance in regression analysis should be questioned in the absence of the appropriate geographic control variables.

[10] See Table D3 in the appendix for control variable coefficient estimates.
[11] See Table D4 in the appendix for control variable coefficient estimates.

Table 3.5: Border-Level Regressions: Overlapping and Multipart Polygon Sensitivity Analysis

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$						
	Sibling Sample Excluding Overlapping Groups		Sibling Sample Excluding Multi-Part Groups		Sibling Sample Excluding Overlapping & Multi-Part Groups	
	(1)	(2)	(3)	(4)	(5)	(6)
CSI Variation (pre-1500)	-0.100**	-0.197***	-0.117***	-0.215**	-0.136***	-0.205*
	(0.041)	(0.073)	(0.045)	(0.086)	(0.050)	(0.106)
Change in CSI Variation (post-1500)	-0.175***	-0.221***	-0.218***	-0.262***	-0.213***	-0.244**
	(0.061)	(0.080)	(0.071)	(0.099)	(0.076)	(0.118)
Controls	No	Yes	No	Yes	No	Yes
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Continental FE	Yes	Yes	Yes	Yes	Yes	Yes
Language Cluster 1	744	744	587	587	492	492
Language Cluster 2	666	666	539	539	453	453
Observations	988	988	763	763	632	632
Adjusted R^2	0.36	0.40	0.36	0.40	0.37	0.39

This table tests the sensitivity of the benchmark estimates by limiting the sibling sample to ethnolinguistic pairs that do not overlap with any other groups and are not composed of multi-part group polygons. Control variables are identical to those in columns (5) and (6) in Table 3.3. The unit of observation is a 100 km buffer zone along the contiguous border segment of each language pair. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

On a more general level, these findings suggest that ethnolinguistic differences are potentially linked to a group's level of state centralization in history. Fenske (2014) shows that the gains from trade resulting from agricultural diversity predicts pre-colonial state centralization, a finding that corroborates the theory of Bates (1983). Because of the long-run persistent effects of pre-colonial state centralization (Gennaioli and Rainer, 2007; Michalopoulos and Papaioannou, 2013), an interesting area of future research would be to document the possible role of the state in the diffusion of culture and language, and how this role interacts with the persistent effect of pre-colonial states on comparative economic development today.

Bibliography

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

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2014). Democracy Does Cause Growth. *NBER Working Paper 20004*.

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

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

Alesina, A., Easterly, W., and Matuszeski, J. (2011). Artificial States. *Journal of the European Economic Association*, 9(2):246–277.

Alesina, A. and Giuliano, P. (2013). Culture and Institutions. *NBER Working Paper 19750*.

Alesina, A. and La Ferrara, E. (2005). Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, 43(3):762–800.

Alesina, A., Michalopoulos, S., and Papaioannou, E. (2016). Ethnic Inequality. *Journal of Political Economy*, 124(2):428–488.

Arriola, L. R. (2009). Patronage and Political Stability in Africa. *Comparative Political Studies*, 42(10):1339–1362.

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

Ashraf, Q., Galor, O., and Klemp, M. (2014). The Out of Africa Hypothesis of Comparative Development Reflected by Nighttime Light Intensity. *MRPA Working Paper 55634*.

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

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.

Baldwin, K. and Huber, J. D. (2010). Economic Versus Cultural Differences: Forms of Ethnic Diversity and Public Goods Provision. *American Political Science Review*, 104(4):644–662.

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

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

Bates, R. H. (1974). Ethnic Competition and Modernization in Contemporary Africa. *Comparative Political Studies*, 6(4):457–484.

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.

Batibo, H. M. (2005). *Language Decline and Death in Africa: Causes, Consequences and Challenges*. Multilingual Matters, Tonawanda.

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

Bloemen, H. G. (2013). Language Proficiency of Migrants: The Relation with Job Satisfaction and Matching. *IZA Discussion Paper 7366*.

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

Bowles, S. and Gintis, H. (2004). Persistent Parochialism: Trust and Exclusion in Ethnic Networks. *Journal of Economic Behavior & Organization*, 55:1–23.

Briggs, R. C. (2014). Aiding and Abetting: Project Aid and Ethnic Politics in Kenya. *World Development*, 64:194–205.

Burgess, R., Miguel, E., Jedwab, R., Morjaria, A., and Padró i Miquel, G. (2015). The Value of Democracy: Evidence from Road Building in Kenya. *American Economic Review*, 105(6):1817–1851.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Caselli, F. and Coleman, W. J. (2013). On the Theory of Ethnic Conflict. *Journal of the European Economic Association*, 11(S1):161–192.

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

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

Collier, P. and Gunning, J. W. (1999). Explaining African Economic Performance. *Journal of Economic Literature*, 37(1):64–111.

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.

Crystal, D. (2003). *English as a Global Language*. Cambridge University Press, Cambridge.

De Luca, G., Hodler, R., Raschky, P. A., and Valsecchi, M. (2015). Ethnic Favoritism: An Axiom of Politics? *CESifo Working Paper 5209*, pages 1–35.

Desmet, K., Breton, M., Ortuño-Ortín, I., and Weber, S. (2011). The Stability and Breakup of Nations: A Quantitative Analysis. *Journal of Economic Growth*, 16(3):183–213.

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

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2015). Culture, Ethnicity and Diversity. *NBER Working Paper 20989*.

Desmet, K., Ortuño-Ortín, I., and Weber, S. (2005). Peripheral Diversity and Redistribution. *CEPR Discussion Papers 5112*.

Desmet, K., Ortuño-Ortín, I., and Weber, S. (2009). Linguistic Diversity and Redistribution. *Journal of the European Economic Association*, 7(6):1291–1318.

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

Dickens, A. (2016a). Ethnolinguistic Favoritism in African Politics. *York University, mimeo*, (August).

Dickens, A. (2016b). Population Relatedness and Cross-Country Idea Flows: Evidence from Book Translations. *York University, mimeo*.

Dreher, A., Fuchs, A., Parks, B. C., Raschky, P. A., and Tierney, M. J. (2015). Aid on Demand: African Leaders and the Geography of China's Foreign Assistance. *CESifo Working Paper 5439*.

Dyen, I., Kruskal, J. B., and Black, P. (1992). An Indoeuropean Classification: A Lexicostatistical Experiment. *Transactions of the American Philosophical Society*, 82(5):1–132.

Easterly, W. and Levine, R. (1997). Africa's Growth Tragedy: Policies and Ethnic Divisions. *The Quarterly Journal of Economics*, 112(4):1203–1250.

Easterly, W. and Levine, R. (2012). The European Origin of Economic Development. *NBER Working Paper 18162*.

Ejdemyr, S., Kramon, E., and Robinson, A. L. (2014). Segregation, Ethnic Favoritism, and the Strategic Targeting of Local Public Goods. *Stanford University, mimeo*.

Englebert, P., Tarango, S., and Carter, M. (2002). Dismemberment and Suffocation: A Contribution to the Debate on African Boundaries. *Comparative Political Studies*, 35(10):1093–1118.

Esteban, J., Mayoral, L., and Ray, D. (2012). Ethnicity and Conflict: An Empirical Study. *American Economic Review*, 102(4):1310–1342.

Esteban, J. and Ray, D. (2011). A Model on Ethnic Conflict. *Journal of the European Economic Association*, 9(3):496–521.

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

Fearon, J. D. and Laitin, D. D. (1999). Weak States, Rough Terrain, and Large-Scale Ethnic Violence Since 1945. *Paper presented at the 1999 Annual Meetings of the American Political Science Association*.

Fenske, J. (2013). Does Land Abundance Explain African Institutions? *The Economic Journal*, 123(573):1363–1390.

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

Franck, R. and Rainer, I. (2012). Does the Leader's Ethnicity Matter? Ethnic Favoritism, Education, and Health in Sub-Saharan Africa. *American Political Science Review*, 106(2):294–325.

Francois, P., Rainer, I., and Trebbi, F. (2015). How Is Power Shared in Africa? *Econometrica*, 83(2):465–503.

Galor, O. and Özak, O. (2015). Land Productivity and Economic Development: Caloric Suitability vs . Agricultural Suitability. *Brown University, mimeo*, pages 1–30.

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

Galor, O., Özak, O., and Sarid, A. (2016). Geographical Origins and Economic Consequences of Language Structures. *Brown University, mimeo*.

Gennaioli, N. and Rainer, I. (2007). The Modern Impact of Precolonial Centralization in Africa. *Journal of Economic Growth*, 12(3):185–234.

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.

Goemans, H. E., Gleditsch, K. S., and Chiozza, G. (2009). Introducting Archigos: A Data Set of Political Leaders. *Journal of Peace Research*, 46(2):269–283.

Gomes, J. F. (2014). The Health Costs of Ethnic Distance: Evidence from Sub-Saharan Africa. pages 1–48.

Greenberg, J. H. (1956). The Measurement of Linguistic Diversity. *Language*, 32(1):109–115.

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

Habyarimana, J., Humphreys, M., Posner, D. N., and Weinstein, J. M. (2009). Coethnicity and Trust. In Cook, K., Levi, M., and Hardin, R., editors, *Whom Can We Trust?*, pages 42–64. Russell Sage Foundation, New York.

Harries, L. (1969). Language Policy in Tanzania. *Journal of the International African Institute*, 39(3):275–280.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring Economic Growth From Outer Space. *The American Economic Review*, 102(2):994–1028.

Herbst, J. (2000). *State and Power in Africa*. Princeton University Press, Princeton.

Hodler, R. and Raschky, P. A. (2014). Regional Favoritism. *The Quarterly Journal of Economics*, 129(2):995–1033.

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.

Huber, J. D. and Suryanarayan, P. (2014). Ethnic Inequality and the Ethnification of Political Parties: Evidence from India. *Columbia University, mimeo*.

Isphording, I. E. and Otten, S. (2013). The Costs of Babylon: Linguistic Distance in Applied Economics. *Review of International Economics*, 21(2):354–369.

Isphording, I. E. and Otten, S. (2014). Linguistic Barriers in the Destination Language Acquisition of Immigrants. *Journal of Economic Behavior and Organization*, 105(5):30–50.

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

Jablonski, R. S. (2014). How Aid Targets Votes: The Impact of Electoral Incentives on Foreign Aid Distribution. *World Politics*, 66(2):293–330.

Jensen, N. and Wantchekon, L. (2004). Resource Wealth and Political Regimes in Africa. *Comparative Political Studies*, 37(7):816–841.

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

Joseph, R. A. (1987). *Democracy and Prebendalism in Nigeria*. Cambridge University Press, Cambridge.

Kasara, K. (2007). Tax Me If You Can: Ethnic Geography, Democracy, and the Taxation of Agriculture in Africa. *The American Political Science Review*, 101(1):159–172.

Kezdi, G. (2004). Robust Standard Error Estimation in Fixed-Effects Panel Models. *Hungarian Statistical Review*, 9:95–116.

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 Lagerlof, N.-P. (2016). Geography and State Fragmentation. *York University, mimeo*.

Kramon, E. and Posner, D. N. (2016). Ethnic Favoritism in Primary Education in Kenya. *Quarterly Journal of Political Science*, 11(1):1–58.

Ku, H. and Zussman, A. (2010). Lingua franca: The role of English in international trade. *Journal of Economic Behavior & Organization*, 75(2):250–260.

Kyriacou, A. P. (2013). Ethnic Group Inequalities and Governance : Evidence from Developing Countries. *Kyklos*, 66(1):78–101.

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

La Porta, R., Lopez-de Silanes, F., Shleifer, A., and Vishny, R. (1999). The Quality of Government. *Journal of Law, Economics, and Organization*, 15(1):222–279.

Leighton, W. A. and López, E. J. (2013). *Madmen, Intellectuals, and Academic Scribblers: The Economic Engine of Political Change*. Stanford University Press, Redwood City.

Lewis, P. M. (2009). *Ethnologue: Languages of the World*. SIL International, Dallas, 16 edition.

Marx, B., Stoker, T. M., and Suri, T. (2015). There is No Free House: Ethnic Patronage in a Kenyan Slum. *MIT, mimeo*.

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

Michalopoulos, S. and Papaioannou, E. (2013). Pre-colonial Ethnic Institutions and Contemporary African Development. *Econometrica*, 81(1):113–152.

Michalopoulos, S. and Papaioannou, E. (2014). National Institutions and Subnational Development in Africa. *The Quarterly Journal of Economics*, 29(1):151–213.

Michalopoulos, S. and Papaioannou, E. (2016). The Long-Run Effects of the Scramble for Africa. *American Economic Review*, 106(7):1802–1848.

Michalopoulos, S., Putterman, L., and Weil, D. N. (2016). The Influence of Ancestral Lifeways on Individual Economic Outcomes in Sub-Saharan Africa. *NBER Working Paper 21907*.

Miguel, E. (2004). Tribe or Nation?: Nation Building and Public Goods in Kenya versus Tanzania. *World Politics*, 56(3):327–362.

Miguel, E. and Gugerty, M. K. (2005). Ethnic Diversity, Social Sanctions and Public Goods in Kenya. *Journal of Public Economics*, 89(11-12):2325–2368.

Mwakikagile, G. (2010). *Ethnic Diversity and Integration in The Gambia: The Land, The People and The Culture*. Continental Press, Dar es Salaam.

Nunn, N. (2008). The Long-Term Effects of Africa's Slave Trades. *The Quarterly Journal of Economics*, 123(1):139–176.

Nunn, N. (2013). Historical Development. In Aghion, P. and Durlauf, S., editors, *Handbook of Economic Growth*, volume 2. North-Holland.

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.

Nunn, N. and Wantchekon, L. (2011). The Slave Trade and the Origins of Mistrust in Africa. *American Economic Review*, 101(7):3221–3252.

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

Oucho, J. (2006). Cross-Border Migration and Regional Initiatives in Managing Migration in Southern Africa. In Kok, P., Gelderblom, D., Oucho, J., and van Zyl, J., editors, *Migration in South and Southern Africa: Dynamics and Determinants*, pages 47–70. HSRC Press, Cape Town.

Padró i Miquel, G. (2007). The Control of Politicians in Divided Societies: The Politics of Fear. *The Review of Economic Studies*, 74(2007):1259–1274.

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

Posner, D. N. (2004). Measuring ethnic fractionalization in Africa. *American Journal of Political Science*, 48(4):849–863.

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

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

Ramankutty, N., Foley, J. A., Norman, J., and McSweeney, K. (2002). A Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Changes. *Global Ecology and Biogeography*, 11:377–392.

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

Solon, G., Haider, S. J., and Wooldridge, J. (2015). What Are We Weighting For? *Journal of Human Resources*, 50(2):301–316.

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

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

Spolaore, E. and Wacziarg, R. (2013b). Long-Term Barriers to Economic Development. *NBER Working Paper 19361*.

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

Spolaore, E. and Wacziarg, R. (2015). Ancestry, Langauge and Culture. *NBER Working Paper 21242*.

Spolaore, E. and Wacziarg, R. (2016). War and Relatedness. *Review of Economics and Statistics*, 98(2):925–939.

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

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

Wantchekon, L. (2003). Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin. *World Politics*, 55(3):399–422.

Wesseling, H. (1996). *Divide and Rule: The Partition of Africa, 1880-1914*. Praeger, Westport.

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

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

Wright, D. R. (2015). *The World and a Very Small Place in Africa: A History of Globalization in Niumi, The Gambia*. Routledge, New York.

Young, A. (2013). Inequality, the Urban-Rural Gap, and Migration. *Quarterly Journal of Economics*, 128(4):1727–1785.

Appendix A

Language Appendix

A.1 Computerized Lexicostatistical Linguistic Distance

The computerized approach to estimating lexicostatistical distances was developed as part of the *Automatic Similarity Judgement Program* (ASJP), a project run by linguists at the Max Planck Institute for Evolutionary Anthropology. To begin a list of 40 implied meanings (i.e., words) are compiled for each language to compare the lexical similarity of any language pair. Swadesh (1952) first introduced the notion of a basic list of words believed to be universal across nearly all world languages. When a word is universal across world languages, its implied meaning, and therefore any estimate of linguistic distance, is independent of culture and geography. From here on I refer to this 40-word list as a Swadesh list, as it is commonly called.[1]

For each language the 40 words are transcribed into a standardized orthography called ASJP-code, 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).[2] 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 $1 \leq i \leq M$.[3] 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

[1] 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).

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

[3] 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.

more edits, the distance is normalized according to word length:

$$LDN(\alpha_i, \beta_i) = \frac{LD(\alpha_i, \beta_i)}{L(\alpha_i, \beta_i)} \tag{A.1}$$

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 between two languages is then

$$LDN(\alpha, \beta) = \frac{1}{M} \sum_{i=1}^{M} LDN(\alpha_i, \beta_i). \tag{A.2}$$

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

$$GD(\alpha, \beta) = \frac{1}{M(M-1)} \sum_{i \neq j}^{M} LD(\alpha_i, \beta_j), \tag{A.3}$$

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 (A.2) with equation (A.3) as follows:

$$LDND(\alpha, \beta) = \frac{LDN(\alpha, \beta)}{GD(\alpha, \beta)}. \tag{A.4}$$

$LDND(\alpha, \beta)$ 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. In Chapter 1, I transform this measure of lexicostatistical distance into a measure of linguistic similarity as follows:

$$LS(\alpha, \beta) = 1 - LDND(\alpha, \beta). \tag{A.5}$$

A.2 Cladistic Similarity

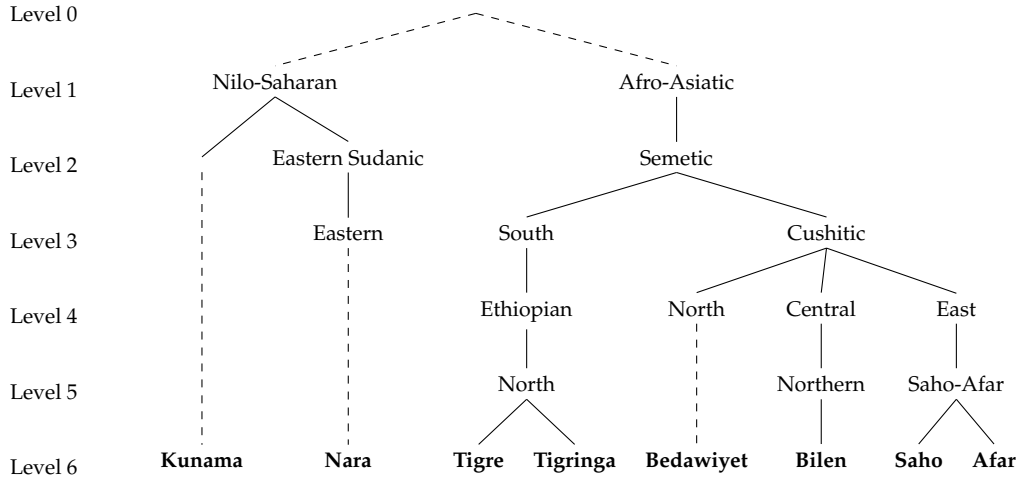
To construct a measure cladistic similarity I calculate the number of shared branches between language α and β on the Ethnologue language tree, denoted $s(\alpha, \beta)$. Let M be the maximum number of tree branches between any two languages. Cladistic similarity is constructed as follows:

$$CS(\alpha, \beta) = \left(\frac{s(\alpha, \beta)}{M} \right)^{\delta}, \tag{A.6}$$

where δ is an arbitrarily assigned weight used to discount more recent linguistic cleavages relative to deep cleavages. This weight is arbitrary because there is no consensus on the assumed value of the weight. Fearon (2003) argues the true function is concave and assumes a value of $\delta = 0.5$, which has since become the convention. Desmet et al. (2009) experiment with a range of values between $\delta \in [0.04, 0.10]$, but settle on a value of $\delta = 0.05$. In all reported estimates I assume $\delta = 0.5$, though the estimates are robust to alternative weighting assumptions.

One issue with calculating cladistic similarity is the asymmetrical nature of historical language splitting. Because the number of branches varies among language families and subfamilies, the maximum number of branches between any two languages is not constant. To overcome this challenge I assume that all current languages are of equal distance from the proto-language at the root of the Ethnologue language tree. I visualize this assumption in Figure A1, where I have constructed a phylogenetic language tree for the 8 distinct languages of Eritrea. The dashed lines represent this assumed historical relationship, so in all cases the contemporary Eritrean languages possess an equal number of branches to the proto-language at Level 0. Although $M = 6$ in Figure A1, in the Ethnologue language tree the highest number of classifications for any language is $M = 15$, which I abstract from here for simplicity.

Figure A1: Phylogenetic Tree of the Eight Major Eritrean Languages



Appendix B

Chapter 1 Appendix

B.1 Data Descriptions, Sources and Summary Statistics

B.1.1 Regional-Level Data Description and Sources

Country-language groups: Geo-referenced country-language group data comes from the World Language Mapping System (WLMS). These data map information from each language in the Ethnologue to the corresponding polygon. When calculating averages within these language group polygons, I use the Africa Albers Equal Area Conic projection.
Source: http://www.worldgeodatasets.com/language/

Linguistic similarity: I construct two measures of linguistic similarity: lexicostatistical similarity from the Automatic Similarity Judgement Program (ASJP), and cladistic similarity using Ethnologue data from the WLMS. I use these to measure the similarity between each language group and the ethnolinguistic identity of that country's national leader. I discuss how I assign a leader's ethnolinguistic identity in Section 1.2.3.
Source: http://asjp.clld.org and http://www.worldgeodatasets.com/language/

Night lights: Night light intensity comes from the Defense Meteorological Satellite Program (DMSP). My measure of night lights is calculated by averaging across pixels that fall within each WLMS country-language group polygon for each year the night light data is available (1992-2013). To minimize area distortions I use the Africa Albers Equal Area Conic projection. In some years data is available for two separate satellites, and in all such cases the correlation between the two is greater than 99% in my sample. To remove choice on the matter I use an average of both. The dependent variable used in the benchmark analysis is $\ln(0.01 + \text{average night lights})$.
Source: http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

Population density: Population density is calculated by averaging across pixels that fall within each country-language group polygon. To minimize area distortions I use the Africa Albers Equal

Area Conic projection. Data comes from the Gridded Population of the World, which is available in 5-year intervals: 1990, 1995, 2000, 2005, 2010. For intermediate years I assume population density is constant; e.g., the 1995 population density is assigned to years 1995-1999. Throughout the regression analysis I use log population density.
Source: http://sedac.ciesin.columbia.edu/data/collection/gpw-v3

National leaders: I collected birthplace locations of all African leaders between 1991-2013. Names of African leaders and years entered and exited office comes from the Archigos Database on Leaders 1875-2004 (Goemans et al., 2009), which I extended to 2011 using data from Dreher et al. (2015), and 2012-2013 using a country's Historical Dictionary and other secondary sources.
Source: http://www.rochester.edu/college/faculty/hgoemans/data.htm

National leader birthplace coordinates: Birthplace locations are confirmed using Wikipedia, and entered into www.latlong.com to collect latitude and longitude coordinates.
Source: http://www.latlong.net

Years in office: To calculate each leader's current years in office and total years in office I use the entry and exit data described above.
Source: Calculated using Stata.

Distance to leader's birth region: Country-language group centroids calculated in ArcGIS, and the distance between each centroid and the national leader's birthplace coordinates is calculated in Stata using the `globdist` command. Throughout the regression analysis I use log leader birthplace distance.
Source: Calculated using ArcGIS and Stata.

Absolute difference in elevation: I collect elevation data from the National Geophysical Data Centre (NGDC) at the National Oceanic and Atmospheric Administration (NOAA). I measure average elevation of each partitioned language group and leader's ethnolinguistic group. To minimize area distortions I use the Africa Albers Equal Area Conic projection. I use Stata to calculate the absolute difference between the two.
Source: www.ngdc.noaa.gov/mgg/topo/globe.html

Absolute difference in ruggedness: As a measure of ruggedness I use the standard deviation of the NGDC elevation data. I use Stata to calculate the absolute difference between the two.
Source: www.ngdc.noaa.gov/mgg/topo/globe.html

Absolute difference in precipitation: Precipitation data comes from the WorldClim – Global Climate Database. I measure average precipitation within each partitioned language group and

leader's ethnolinguistic group using the Africa Albers Equal Area Conic projection. I use Stata to calculate the absolute difference between the two.
Source: http://www.worldclim.org/current

Absolute difference in temperature: Temperature data comes from the WorldClim – Global Climate Database. I measure the average temperature within each partitioned language group and leader's ethnolinguistic group using the Africa Albers Equal Area Conic projection. I use Stata to calculate the absolute difference between the two.
Source: http://www.worldclim.org/current

Absolute difference in caloric suitability index: I sourced the caloric suitability index (CSI) data from Galor and Özak (2016). CSI is a measure of agricultural productivity that reflects the caloric potential in 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 the pre-1500 average CSI measure that includes cells with zero productivity. The results are not sensitive to which measure I use. I measure average CSI within each partitioned language group and leader's ethnolinguistic group using the Africa Albers Equal Area Conic projection. I use Stata to calculate the absolute difference between the two.
Source: http://omerozak.com/csi

Oil reserve: I construct an indicator variable equal to one if an oil field is found in both the partitioned language group and leader's ethnolinguistic group. Version 1.2 of the Petroleum Dataset contains geo-referenced point data indicating the presence of on-shore oil and gas deposits from around the world.
Source: https://www.prio.org/Data/Geographical-and-Resource-Datasets/Petroleum-Dataset/

Diamond reserve: I construct an indicator variable equal to one if a known diamond deposit is found in both the partitioned language group and leader's ethnolinguistic group. Version 1.2 of the Petroleum Dataset contains geo-referenced point data indicating the presence of on-shore oil and gas deposits from around the world.
Source: https://www.prio.org/Data/Geographical-and-Resource-Datasets/Diamond-Resources/

B.1.2 Individual-Level Data Description and Sources

Unless otherwise stated, all individual-level data comes from the Demographic and Health Surveys (DHS). Source: http://dhsprogram.com/

Individual linguistic similarity: To assign an individual a home language I assign the reported language a respondent speaks at home when this data is available (59 percent availability). For surveys when this data isn't available or the reported language is "other", I map the respondent's

home language from their reported ethnicity. To do this I use the following assignment rule:

1. Direct match: the DHS ethnicity name is the same as an Ethnologue language name for the respondent's country of residence.
2. Alternative name: the unmatched DHS ethnicity is an unambiguous alternative name for a language in the Ethnologue or Glottolog database.
3. Macrolanguage: if the ethnicity corresponds to a macrolanguage in the Ethnologue, then I assign the most populated sub-language of that macrolanguage.
4. Population size: if the unmatched ethnicity maps to numerous languages, I choose the language with the largest Ethnologue population.

I also cross-reference the Wikipedia page for each ethnic group to corroborate that the assigned language maps into the reported ethnicity. Then using the same data on leaders as in the regional-analysis, I match the lexicostatistical similarity of the respondent's home language to the leader's ethnolinguistic identity.
Source: http://asjp.clld.org

Locational linguistic similarity: I project DHS cluster latitude and longitude coordinates onto the Ethnologue language map and assign the associated language as the regional language group to that respondent. In instances of overlapping language groups, I assign the largest group in terms of population. Then using the same data on leaders as in the regional-analysis, I match the lexicostatistical similarity of the respondent's home language to the leader's ethnolinguistic identity.
Source: http://asjp.clld.org

Wealth Index: I use the quantile DHS wealth index. The quantile index is derived from a composite measure of a household's assets (e.g., television, refrigerator, telephone, etc.) and access to public resources (e.g., water, electricity, sanitation facility, etc.), in addition to data indicating if a household owns agricultural land and if they employ a domestic servant. Principal component analysis is used to construct the original index, then respondents are order by score and sorted into quintiles. Read the DHS Comparative Report: The DHS Wealth Index for more details.

Age: Age of respondent at the time of survey.

Gender: An indicator variable equal to one if a respondent is female.

Rural: An indicator variable for rural locations.

Education: The 10 education fixed effects are from question 90.

Religion: The 18 fixed effects for the religion of a respondent come from question 91.

Distance to the capital: I use the World Cities layer available on the ArcGIS website, which includes latitude-longitude coordinates and indicators for capital cities. I calculate language group centroids coordinates using ArcGIS, and measure the geodesic distance between the two points in Stata using the `globdist` command.
Source: http://www.arcgis.com/home/

Distance to the coast: I use the coastline shapefile from Natural Earth, calculate the nearest coastline from a language groups centroid using the Near tool in ArcGIS. I measure the geodesic distance between the two points in Stata using the `globdist` command.
Source: http://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-coastline/

Distance to the border: I use country boundaries from the Digital Chart of the World (5^{th} edition) that's complimentary to the Ethnologue data from the WLMS, and calculate the nearest border from a language groups centroid using the Near tool in ArcGIS. I measure the geodesic distance between the two points in Stata using the `globdist` command.
Source: http://www.worldgeodatasets.com/language/

B.1.3 Summary Statistics and Additional Details

Table B1: Language Groups Included in Regional-Level Analysis

Sample	Language Groups
Regional-Level Analysis	Acholi, Adamawa Fulfulde, Adele, Afade, Afrikaans, Alur, Anuak, Anufo, Anyin, Baatonum, Badyara, Baka, Bari, Bata, Bayot, Bedawiyet, Bemba, Berta, Bissa, Boko, Bokyi, Bomwali, Borana-Arsi-Guji Oromo, Buduma, Central Kanuri, Chadian Arabic, Chidigo, Cokwe, Daasanach, Dan, Dazaga, Dendi, Dholuo, Diriku, Ditammari, Ejagham, Ewe, Fur, Gbanziri, Gidar, Glavda, Gola, Gourmanchema, Gude, Gumuz, Hausa, Herero, Holu, Jola-Fonyi, Juhoan, Jukun Takum, Jula, Kaba, Kacipo-Balesi, Kako, Kakwa, Kalanga, Kaliko, Kaonde, Kasem, Khwe, Kikongo, Kisikongo, Kiswahili, Komo, Konkomba, Koromfe, Kuhane, Kunama, Kunda, Kuo, Kuranko, Kusaal, Kwangali, Kxauein, Langbashe, Lozi, Lugbara, Lunda, Lutos, Luvale, Maasai, Madi, Makonde, Mambwe-Lungu, Mandinka, Mandjak, Manga Kanuri, Mann, Manyika, Masana, Mashi, Mbandja, Mbay, Mbukushu, Mende, Monzombo, Moore, Mpiemo, Mundang, Mundu, Musey, Musgu, Nalu, Naro, Ndali, Ndau, Ngangam, Ngbaka Mabo, Ninkare, Northern Kissi, Northwest Gbaya, Nsenga, Ntcham, Nuer, Nyakyusa-Ngonde, Nyanja, Nzakambay, Nzanyi, Nzema, Oshiwambo, Pana, Peve, Pokoot, Psikye, Pulaar, Pular, Runga, Rwanda, Saho, Shona, Shuwa Arabic, Somali, Soninke, Southern Birifor, Southern Kisi, Southern Sotho, Susu, Swati, Taabwa, Talinga-Bwisi, Tamajaq, Tedaga, Teso, Tigrigna, Tonga, Tswana, Tumbuka, Tupuri, Vai, Venda, Wandala, Western Maninkakan, Xhosa, Xoo, Yaka, Yaka, Yalunka, Yao, Yeyi, Zaghawa, Zande, Zarma, Zemba, Zulu

Table B2: Language Groups Included in DHS Individual-Level Analysis

Sample	Language Groups
Individual-Level Analysis (Locational)	Alur, Bemba, Borana, Kaonde, Kasem, Kisi (Southern), Kissi (Northern), Kuhane, Kuranko, Lamba, Lugbara, Lunda, Maninkakan (Western), Mann, Oromo (Borana-Arsi-Guji), Pular, Somali, Soninke, Susu, Taabwa, Teso
Individual-Level Analysis (Individual)	Afar, Amharic, Aushi, Bamanankan, Bandi, Bemba, Berta, Bissa, Bobo Madare (Southern), Bwile, Cokwe, Dagaare (Southern), Dagbani, Dan, Dholuo, Ekegusii, Farefare, Ganda, Gedeo, Gikuyu, Gola, Gourmanchema, Gwere, Hadiyya, Harari, Hausa, Ila, Jola-Fonyi, Kamba, Kambaata, Kaonde, Kigiryama, Kipsigis, Kisi (Southern), Kissi (Northern), Kono, Koongo, Kpelle (Guinea), Kpelle (Liberia), Krio, Kuhane, Kunda, Kuranko, Lala-Bisa, Lamba, Lendu, Lenje, Limba (East), Lozi, Luba-Kasai, Lugbara, Lunda, Luvale, Maasai, Madi, Mambwe-Lungu, Mandinka, Maninkakan (Kita), Mann, Mbunda, Mende, Moore, Ngombe, Nkoya, Nsenga, Nyanja, Oromo (Borana-Arsi-Guji), Oromo (West Central), Oyda, Pulaar, Pular, Rendille, Samburu, Sebat Bet Gurage, Senoufo (Mamara), Serer-Sine, Sherbro, Sidamo, Soli, Somali, Songhay (Koyra Chiini), Soninke, Susu, Swahili, Taabwa, Tamasheq, Teso, Themne, Tigrigna, Tonga, Tumbuka, Turkana, Wolaytta, Wolof

Table B3: Leaders Included in Regional-Level Analysis

Sample	Leaders
Regional-Level Analysis	**Angola:** José Eduardo dos Santos; **Benin:** Thomas Yayi Boni, Mathieu Kérékou; **Botswana:** Quett Masire, Festus Mogae; **Burkina Faso:** Blaise Compaoré; **Cameroon:** Paul Biya; **Central African Republic:** Ange-Félix Patassé, André-Dieudonné Kolingba; **Chad:** Idriss Déby; **Congo:** Pascal Lissouba, Denis Sassou Nguesso; **Côte d'Ivoire:** Konan Bedie, Laurent Gbagbo, Robert Guéï, Félix Houphouët-Boigny, Alassane Ouattara; **DRC:** Joseph Kabila, Laurent-Désiré Kabila, Mobutu Sese Seko; **Eritrea:** Isaias Afewerki; **Ethiopia:** Hailemariam Desalegn, Meles Zenawi; **Gambia:** Yahya Jammeh, Dawda Jawara; **Ghana:** John Evans Atta-Mills, John Agyekum Kufuor, John Dramani Mahama, Jerry Rawlings; **Guinea:** Moussa Dadis Camara, Alpha Condé, Lansana Conté, Sékouba Konaté; **Guinea-Bissau:** Kumba Ialá, Manuel Serifo Nhamadjo, Henrique Periera Rosa, Malam Bacai Sanhé, João Bernardo Vieira; **Kenya:** Daniel arap Moi; Mwai Kibaki; **Lesotho:** Elias Phisoana Ramaema, Ntsu Mokhehle, Pakalithal Mosisili, Tom Thabane; **Liberia:** Gyude Bryant, Ruth Perry, Wilton G. S. Sankawulo, Ellen Johnson Sirleaf, Charles Taylor; **Malawi:** Hastings Kamuzu Banda, Joyce Banda, Bakili Muluzi, Bungu wa Mutharika; **Mali:** Alpha Oumar Konaré, Amadou Toumani Touré, Dioncounda Traoré; **Mozambique:** Armando Guebuza, Joaquim Chissano; **Namibia:** Sam Nujoma, Hifikepunye Pohamba; **Niger:** Mahamadou Issoufou, Ibrahim Baré Maïnassara, Mahamane Ousmane, Ali Saibou, Mamadou Tandja; **Nigeria:** Sani Abacha, Abdulsalami Abubakar, Goodluck Jonathan, Olusegun Obasanjo, Umaru Musa Yar'Adua; **Senegal:** Abdou Diouf, Macky Sall, Abdoulaye Wade; **Sierra Leone:** Ahmad Tejan Kabbah, Ernest Bai Koroma, Johnny Paul Koroma, Valentine Strasser; **Somalia:** Abdullahi Yusuf Ahmed, Sharif Sheikh Ahmed, Abdiqasim Salad Hassan, Hassan Sheikh Mohamud, Ali Mahdi Muhammad; **South Africa:** Frederik Willem de Klerk, Nelson Mandela, Thabo Mbeki, Jacob Zuma; **Sudan:** Omar Hassan Ahmad al-Bashir; **Tanzania:** Jakaya Kikwete, Benjamin Mkapa, Ali Hassan Mwinyi; **Togo:** Gnassingbé Eyadéma, Faure Gnassingbé; **Uganda:** Yoweri Museveni; **Zambia:** Frederick Chiluba, Levy Mwanawasa, Michael Sata; **Zimbabwe:** Robert Mugabe

Table B4: Leaders Included in DHS Individual-Level Analysis

Sample	Leaders
Individual-Level Analysis	**Burkina Faso:** Blaise Compaoré **Democratic Republic of Congo:** Joseph Kabila **Ethiopia:** Meles Zenawi **Ghana:** Jerry Rawlings; John Agyekum Kufuor **Guinea:** Alpha Condé; Lansana Conté **Kenya:** Mwai Kibaki **Liberia:** Ellen Johnson Sirleaf **Mali:** Alpha Oumar Konaré; Amadou Toumani Touré **Namibia:** Hifikepunye Pohamba **Senegal:** Abdou Diouf; Abdoulaye Wade **Sierra Leone:** Ernest Bai Koroma **Uganda:** Yoweri Museveni **Zambia:** Levy Mwanawasa; Michael Sata

Table B5: Countries Included in Regional- and Individual-Level Analysis

Sample	Countries
Regional-Level Analysis	Angola, Benin, Botswana, Burkina Faso, Cameroon, Central African Republic, Chad, Congo, Cote d'Ivoire, Democratic Republic of Congo, Eritrea, Ethiopia, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Malawi, Mali, Mozambique, Namibia, Niger, Nigeria, Senegal, Sierra Leone, Somalia, South Africa, Sudan, Tanzania, Togo, Uganda, Zambia, Zimbabwe
Individual-Level Analysis	Burkina Faso, Democratic Republic of Congo, Ethiopia, Ghana, Guinea, Kenya, Liberia, Mali, Namibia, Senegal, Sierra Leone, Uganda, Zambia

Table B6: Summary Statistics – Regional-Level Dataset

	Mean	Std dev.	Min	Max	N
Night lights$_t$	0.123	0.387	0.000	4.540	6,610
ln(0.01 + night lights$_t$)	-3.487	1.427	-4.605	1.515	6,610
ln(0.01 + night lights$_{t-1}$)	-3.507	1.415	-4.605	1.515	6,315
$\sqrt{\text{night lights}_t}$	0.187	0.297	0.000	2.131	6,610
ln(night lights$_t$)	-3.370	2.049	-10.60	1.513	4,069
Lexicostatistical similarity$_{t-1}$	0.193	0.230	0.000	1.000	6,610
Cladistic similarity$_{t-1}$	0.409	0.330	0.000	1.000	6,610
Coethnicity$_{t-1}$	0.047	0.212	0.000	1.000	6,610
Non-coethnic cladistic similarity$_{t-1}$	0.362	0.313	0.000	0.966	6,610
Non-coethnic lexicostatistical similarity$_{t-1}$	0.146	0.148	0.000	0.960	6,610
Lexicostatistical similarity$_{t+1}$	0.194	0.230	0.000	1.00	6228
Current years in office$_{t-1}$	11.44	8.680	1.000	38.00	6,610
Total years in office$_{t-1}$	18.50	10.19	1.000	38.00	6,610
Log distance (km) to leader's group$_{t-1}$	5.844	1.485	0.000	7.419	6,610
Log population density$_t$	2.886	1.529	-2.169	6.116	6,610
Absolute difference in elevation$_t$	250.5	296.1	0.000	2,021	6,610
Absolute difference in ruggedness$_t$	101.5	105.5	0.000	542.4	6,610
Absolute difference in precipitation$_t$	30.20	28.90	0.00	230.7	6,610
Absolute difference in mean temperature$_t$	16.81	17.09	0.000	120.2	6,610
Absolute difference in caloric suitability index$_t$	298.0	310.1	0.000	1711	6,610
Oil reserve in both leader and language group$_t$	0.018	0.131	0.000	1.000	6,610
Diamond mine in both leader and language group$_t$	0.079	0.269	0.000	1.000	6,610
Absolute difference in malaria suitability$_t$	4.951	5.635	0.000	29.30	5,111
Absolute difference in land suitability$_t$	0.178	0.184	0	0.777	5111
Democracy$_{t-1}$	0.435	4.877	-9.000	9.000	6,573
Language group population share	0.045	0.113	0	0.851	6610
Distance (km) to capital city	559.7	397.7	26.58	1922	6,610
Distance (km) to the coast	677.9	408.4	10.52	1743	6,610

Table B7: Summary Statistics – DHS Individual-Level Dataset

	Mean	Std Dev.	Min	Max	N
Wealth index	2.974	1.468	1.000	5.000	56,455
Locational similarity	0.350	0.380	0.025	1.000	56,455
Individual similarity	0.363	0.387	0.021	1.000	56,455
Age	29.36	10.51	15.00	78.00	56,455
Female indicator	0.663	0.473	0.000	1.000	56,455
Rural indicator	0.635	0.482	0.000	1.000	56,455
Education	4.721	1.520	1.000	6.000	56,455
Religion	4.912	2.032	1.000	8.000	56,455
Log distance to the coast (km)	6.059	0.910	1.654	7.238	56,455
Log distance to the border (km)	4.948	0.887	0.920	6.801	56,455
Log distance to the capital (km)	5.676	0.727	2.070	7.548	56,455

Table B8: Summary Statistics – Power Sharing Dataset

	Mean	Std dev.	Min	Max	N
Share of cabinet positions$_t$	0.056	0.078	0.000	0.471	2,539
Share of top cabinet positions$_t$	0.057	0.108	0.000	0.643	2,539
Share of low cabinet positions$_t$	0.055	0.078	0.000	0.450	2,539
Coethnicity$_t$	0.077	0.266	0.000	1.000	2,539
Lexicostatistical similarity$_t$	0.196	0.267	0.000	1.000	2,539
Non-coethnic lexicostatistical similarity$_t$	0.114	0.122	0.000	0.659	2,539
Ethnic group population share$_t$	0.057	0.065	0.005	0.390	2,539

B.2 Mapping Ethnicity to Language

There is mostly agreement between ethnographers that language is a suitable marker of ethnicity in Africa (Batibo, 2005; Desmet et al., 2015). The challenge of mapping ethnicity to language is that, in some instances, a single ethnic group speaks many languages. In such instances it's not obvious what language is the appropriate language to match to a leader's ethnicity. As a solution to this problem I use the following three-step assignment rule to construct a mapping between ethnicity and language in Africa.

Step 1: For each ethnic group, I refer to the Ethnologue list of languages for the country to which they belong. If a language name is identical to the ethnic name then I assign the corresponding language to that ethnicity.

Step 2: If there is no language name identical to the ethnicity then I check the alternate names for a language. If an ethnic name matches an alternate language name, I assign the corresponding language to that ethnicity.

Step 3: If a set of potential language matches still exist, I assign the largest language group (in terms of population) to the ethnic group.

B.3 Supplementary Materials

This section presents results referenced but not presented in the main body of the paper.

B.3.1 Various Fixed Effects Specifications

Table B9 reports 27 different estimates: 9 versions of equation (1.1) for each of the 3 linguistic similarity measures. Columns 1-3 report between-group estimates with country-year fixed effects, the estimates in columns 4-6 add country-language fixed effects, and columns 7-9 report estimates for the triple-difference estimator. For each set of three regressions I report estimates (i) without any covariates, (ii) estimates that only control for log population density and the logged geodesic distance between each partitioned group and the corresponding leader's group, and (iii) the full set of covariates I outlined in Section 1.3.

Consistent with my hypothesis of ethnolinguistic favoritism, all 27 coefficients are positive and the majority are statistically significant. In all cases my preferred measure of lexicostatistical similarity is significant with the exception of column 4, where lexicostatistical similarity has a reported p-value of 0.127. However, in this instance, the estimator lacks language-year fixed effects and thus does not exploit the counterfactual comparison of the same language group on the other side of the border.

Indeed, the addition of language-year fixed effects in 7-9 adds considerable precision to the estimates relative to columns 4-6. The allowance of a within-group estimator that comes from having a panel of partitioned language groups substantially improves my ability to identify ethnolinguistic favoritism.

I also provide estimates for cladistic similarity and coethnicity to see how these alternative measures compare to lexicostatistical similarity. For my benchmark estimates both coefficients are positive and statistically significant, albeit only at the 10 percent level. Not only does the estimated coefficient monotonically increase in the measured continuity of linguistic similarity, but lexicostatistical similarity is also more precisely estimated than both alternative measures. This suggests that the observable variation among non-coethnic groups assists in identifying patterns of ethnic favoritism in Africa.

B.3.2 Additional Controls

In this section I reproduce the benchmark estimates with two additional control variables: the Malaria Ecology Index (Kiszewski et al., 2004) and the Agricultural Suitability Index (Ramankutty et al., 2002). The trouble with these data is that in a number of instances a single raster cell covers an area larger than a country-language group partition because these data are only available at a spatial resolution of $0.5^\circ \times 0.5^\circ$ (approximately 111 km $\times$ 111 km). These partitions are dropped from group average calculations, resulting in a sample 61.5 percent of the benchmark sample size.

Table B10 reports these subsample estimates that include the additional control variables. For each of the three measures of similarity I report estimates that include the absolute difference in the

Malaria Ecology Index, the absolute difference in the Agricultural Suitability Index and estimates that include both measures, in addition to benchmark set of controls. The results are unchanged by including these controls.

B.3.3 Measurement Error

When an unambiguous assignment of a leader's ethnolinguistic identity cannot be made, I assign the group with the largest population among the set of potential matches. The finding that favoritism exists among groups that are not coethnic to the leader might be driven by the measurement error introduced by this approach.

In this section I report estimates on a subsample of my benchmark dataset that excludes the 4 leaders I could not unambiguously match.[1] Table B11 reports these results. Overall little is changed from my benchmark estimates, with the exception that coethnicity is no longer significant at standard levels of confidence. However, lexicostatistical similarity is roust to these excluded leaders, and most importantly, column (4) of Table B11 makes clear that the significance of non-coethnic similarity is not a consequence of the possible measurement error introduced when assigning an ethnolinguistic identity to the aforementioned leaders.

B.3.4 Balanced Panel

In this section I test the robustness of the benchmark estimates using a balanced panel of country-language groups between 1992 and 2013. My benchmark panel was unbalanced because of missing data on language lists used to estimate lexciostatistical similarity. This is problematic if these lists are missing for non-random reasons (Cameron and Trivedi, 2005). To check this I limit the analysis to a balanced sample of 84 language groups partitioned across 23 countries. Table B12 reports these estimates.

In all 27 reported regressions the measure of linguistic similarity takes the expected positive sign positive. For my preferred measure of lexicostatistical similarity the coefficients are statistically significant in all but one regression. The magnitudes of the estimates are also relatively similar to my benchmark estimates. To the contrary cladistic similarity seems to be quite sensitive to this subsample and in only significant in a single instance. The coethnic results are similar to those in Table 1.3.

B.3.5 Weighted Regressions

In this section I test for heteroskedasticity in my benchmark estimates by weighting regressions by the Ethnologue population of each language group. The idea is that the measure of night light intensity is an average within each country-language group, and it is likely to have more variance in places where the population is small (Solon et al., 2015). Table B13 reports these estimates.

[1] Mobutu Sese Seko (DRC), Joseph Kabila (DRC), Laurent-Desire Kabila (DRC) and Goodluck Jonathan (Nigeria).

The lexicostatistical estimates are less sensitive to weighting than the cladistic and coethnic estimates. While a few lexicostatistical estimates lose their significance in columns (4)-(6), these estimates do not exploit language-year fixed effects, and hence are not identified off the exogenous within-group variation. In my benchmark specification in column (9), the effect of lexicostatistical similarity is significant at the 5 percent level and very similar to the benchmark estimate in terms of magnitude.

B.3.6 Alternative Night Light Transformations

The log transformation used throughout the regional analysis is without a doubt arbitrary. The use of this transformation has become the convention when using these night lights data so I follow the literature in my chose to add 0.01 to the log transformation. Nonetheless, I experiment with two alternative transformations in Table B14.

In columns (1)-(3) I report estimates where the dependent variable is defined as the square root of the raw night lights data. In columns (4)-(6) I log the night lights data without adding a constant. The latter results in a substantial loss of observations due to the fact that 40 percent of the observations exhibit zero night light activity. Because I must observe a partitioned group on both sides of the border for any year, I lose nearly 60 percent of my benchmark sample using this log transformation.

I find that the lexicostatistical estimate is robust to both transformations, while the cladistic is only robust to the square root transformation. Coethnicity remains positive but loses its statistical significance is both instances.

B.3.7 DHS Additional Tables

Table B17 reports 15 estimates: 5 separate specifications for both locational and individual similarity, and the same five specifications for the joint similarity estimates. In all specifications I adjust standard errors for clustering in country-wave-locational-language areas.

The top panel reports estimates for locational similarity. In column (1) the coefficient takes the expected positive sign, but is insignificant because the standard error is estimated to be quite large. However, in this specification I do not account for any individual characteristics, including whether a respondent lives in a rural location. Young (2013) shows that the urban-rural income gap accounts for 40 percent of mean country inequality in a sample of 65 DHS countries. In column (2) I report an estimate that includes a rural indicator variable. Indeed, the inclusion of this indicator substantially improves the precision of estimation, where locational similarity is now significant at the 1 percent level. In column (3) I add a set of individual controls.[2] The magnitude of locational similarity increases slightly and maintains its strong significant effect on individual wealth. In Table B17 I add each individual control variable one at a time. While I

[2]The set of individual controls include age, age squared, a female indicator, a rural indicator, a capital city indicator, 5 education fixed effects and 7 religion fixed effects. See Appendix B for variable definitions.

account for capital city effects with an indicator variable, I also account for additional spatial effects in columns (4) and (5) by separately adding the geodesic distance to the nearest coast and border.[3]

The middle panel of Table B15 reports estimates for individual similarity. While all coefficients take the expected positive sign, only a single estimate of individual similarity is statistically significant. When I do not control for any covariates the effect of individual similarity is very precisely estimated. To the contrary, the effect goes away once I account for respondents living in rural locations. The same is true when including the full set of controls.

Next I jointly estimate both channels using the aforementioned variation among individuals non-native to the region in which they reside. The results are consistent with the rest of the table and reported in the bottom panel of Table B15. In column (1) the estimate for individual similarity outperforms locational similarity when no individual characteristics are accounted for, however the reverse is true in columns (2)-(5) as covariates are incrementally added – in particular the rural indicator.

To show that the locational mechanism is not only driven by the coethnic effect, I separately estimate locational coethnicity and non-coethnic locational similarity. I do this in the same way I did in the regional-level analysis: I define non-coethnic locational similarity as $(1 - \text{coethnicity}) \times$ locational similarity. Table B16 reports these estimates. While non-coethnic locational similarity is estimated to be no different than zero in the most basic regression, once again after the baseline set of controls are added both the coethnic and non-coethnic effect are positive and strongly significant. Using the more conservative estimates of column (5), this suggests that the average level of non-coethnic locational similarity (0.164) yields an increase of 0.094 ($= 0.164 \times 0.573$) in the wealth index – roughly one fourth the coethnic effect.

Finally, I also report the DHS estimates for locational similarity and include each baseline covariate one at a time. The idea here is to highlight the relative importance of controlling for the urban-rural inequality gap when using the DHS wealth index (Young, 2013). Table B17 reports these estimates.

Indeed I find that the precision of the locational similarity estimate is substantially improved by including an indicator variable for respondents living in rural regions. While many of the other covariates are themselves positive, no other variable have such a large confounding effect on locational similarity in its absence.

[3] I include distances separately because language areas tend to be fairly small, so location clusters in a partition are usually very close together and distance measures are highly collinear.

B.3.8 Coalition Building

Data

I use data from Francois et al. (2015) on the share of an ethnic group's representation in the governing coalition for 15 African countries.[4] These data are available at a yearly interval until 2004 for the ethnic groups listed in Alesina et al. (2003) and Fearon (2003). Because the unit of observation is an ethnic group, I assign an Ethnologue language group to each ethnicity using the assignment strategy outlined in Appendix B.[5] I measure the lexicostatistical similarity of these groups to the ethnolinguistic identity of the national leader between 1992 and 2004 using the leader data described in Section 1.2.3. In each country a residual ethnic categorization named Other is assigned to capture all groups outside of a country's major ethnic groups. Because Others lack a single ethnolinguistic identity, I assign Other groups a value of zero percent similarity to their leader.

Results

I report estimates of equation (1.2) in Table B18. Column 1 replicates the main estimate of Francois et al. (2015) on the subset of data that I observe lexicostatistical similarity. The coefficient for coethnicity takes the expected positive sign, implying there is a 9 percent increase in the leader's group share of the governing coalition over and above the ministerial appointments made in accordance with the leader's group size. The magnitude of this coefficient is slightly smaller than the comparable coefficient in Francois et al.'s (2015) Table III. This suggests that, if anything, this subsample biases the coefficient downward. Column 2 corroborates this result using lexicostatistical similarity in place of coethnicity. In column 3, I separate the effect of coethnicity from lexicostatistical similarity using the same approach I used in Section 1.4; i.e., non-coethnic similarity $= (1 - \text{coethnicity}) \times$ lexicostatistical similarity. The reported estimates in column 3 confirm that linguistic similarity predicts a group's representation in the governing coalition even among non-coethnic groups.

In columns 4-6 I explore the allocation of top positions in the governing coalition, and in columns 7-9 the allocation of positions outside of the top.[6] In all cases the variables of interest are positive and statistically significant. The most notable observation in this table is remarkable consistency in the magnitude of non-coethnic similarity across specifications. Related groups outside of the leader's ethnic group benefit from receiving positions both low and high in the hierarchy of government.[7]

[4] Benin, Cameroon, Cote dIvoire, Democratic Republic of Congo, Gabon, Ghana, Guinea, Liberia, Nigeria, Republic of Congo, Sierra Leone, Tanzania, Togo, Kenya, and Uganda.

[5] For 87.5 percent of the 264 ethnic groups not listed as "Other", the name of the ethnic group unambiguously corresponds to an Ethnologue name or alternative name in the country in which the group resides. Only 12.5 percent of groups require I use population as a tie breaker when multiple languages can be mapped to an ethnicity. 51 of the assigned languages do not possess an ASJP language list and thus are dropped from the analysis.

[6] Top positions include the president and deputies, as well as ministers of defence, budget, commerce, finance, treasury, economy, agriculture, justice, and state/foreign affairs.

[7] The estimates for group size are statistically significant in all instances. The estimates are also comparable in magnitude to those in Table 3 of Francois et al. (2015), and similarity show evidence of concavity in group size.

Table B9: Benchmark Regressions Using Various Combinations of Fixed Effects

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	1.292***	0.806***	0.936***	0.115	0.200**	0.213**	0.244**	0.297**	0.305***
	(0.255)	(0.306)	(0.318)	(0.075)	(0.087)	(0.088)	(0.112)	(0.120)	(0.116)
Adjusted R^2	0.342	0.428	0.452	0.921	0.921	0.922	0.925	0.925	0.926
Cladistic similarity$_{t-1}$	0.835***	0.488**	0.446**	0.044	0.065	0.058	0.221**	0.219**	0.185*
	(0.199)	(0.205)	(0.203)	(0.064)	(0.066)	(0.068)	(0.104)	(0.102)	(0.103)
Adjusted R^2	0.331	0.428	0.449	0.921	0.921	0.921	0.925	0.925	0.925
Coethnic$_{t-1}$	1.058***	0.386	0.648**	0.092	0.193**	0.202**	0.130	0.139	0.168*
	(0.244)	(0.325)	(0.314)	(0.064)	(0.084)	(0.082)	(0.099)	(0.098)	(0.094)
Adjusted R^2	0.332	0.423	0.447	0.921	0.921	0.922	0.925	0.925	0.925
Geographic controls	No	No	Yes	No	No	Yes	No	No	Yes
Distance & population density	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Language-year fixed effects	No	No	No	No	No	No	Yes	Yes	Yes
Country-language fixed effects	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355	355	355	355
Countries	35	35	35	35	35	35	35	35	35
Language groups	163	163	163	163	163	163	163	163	163
Observations	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610

This table reports benchmark estimates associating each measure of linguistic similarity with night light luminosity for the years $t = 1992 - 2013$. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t - 1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Distance & population density measure the log distance between each country-language group and the leader's ethnolinguistic group, and the log population density of a country-language group, respectively. The geographic controls include the absolute difference in elevation, ruggedness, precipitation, temperature and the caloric suitability index between leader and country-language group regions, in addition to two dummy variables indicating if either region contains diamond and oil deposits. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table B10: Robustness Check: Benchmark Regressions with Additional Control Variables

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.384***	0.368***	0.380***						
	(0.120)	(0.122)	(0.120)						
Cladistic similarity$_{t-1}$				0.255**	0.242**	0.256**			
				(0.114)	(0.111)	(0.111)			
Coethnic$_{t-1}$							0.271**	0.257**	0.269**
							(0.108)	(0.109)	(0.109)
Malaria control	Yes	No	Yes	Yes	No	Yes	Yes	No	Yes
Land suitability control	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Geographic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	228	228	228	228	228	228	228	228	228
Countries	33	33	33	33	33	33	33	33	33
Language groups	105	105	105	105	105	105	105	105	105
Adjusted R^2	0.950	0.949	0.950	0.949	0.949	0.949	0.949	0.949	0.949
Observations	4,065	4,065	4,065	4,065	4,065	4,065	4,065	4,065	4,065

This table reports estimates associating each measure of linguistic similarity with night light luminosity for the years $t = 1992 - 2013$. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t - 1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. Distance & population density measure the log distance between each country-language group and the leader's ethnolinguistic group, and the log population density of a country-language group, respectively. The geographic controls include the absolute difference in elevation, ruggedness, precipitation, temperature and the caloric suitability index between leader and country-language group regions, in addition to two dummy variables indicating if either region contains diamond and oil deposits. The malaria controls measures the absolute difference in the Malaria Ecology Index between leader and country-language groups, while the land suitability control measures the absolute difference in Ramankutty et al.'s (2002) Agricultural Suitability Index. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table B11: Robustness Check: Excluding Leaders with Ambiguous Ethnolinguistic Identities

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$					
	(1)	(2)	(3)	(4)	(5)
Lexicostatistical similarity$_{t-1}$	0.278**				
	(0.116)				
Cladistic similarity$_{t-1}$		0.199*			
		(0.108)			
Coethnicity$_{t-1}$			0.145	0.229**	0.218*
			(0.095)	(0.104)	(0.112)
Non-coethnic lexicostatistical similarity$_{t-1}$				0.480**	
				(0.237)	
Non-coethnic cladistic similarity$_{t-1}$					0.185
					(0.130)
Geographic controls	Yes	Yes	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes
Clusters	314	314	314	314	314
Countries	34	34	34	34	34
Language groups	144	144	144	144	144
Adjusted R^2	0.922	0.922	0.922	0.922	0.922
Observations	5,745	5,745	5,745	5,745	5,745

This table reports estimates from a subsample that excludes all ambiguous leadership assignments. Because these problematic assignments introduce measurement error, excluding them from the analysis ensures that the results are not a consequence of measurement. Average night light intensity is measured in language group l of country c in year t, and Lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the national leader and is measured on the unit interval. The same log transformation of the dependent variable is used for the lagged value of night lights, i.e., $\ln(0.01 + NightLights_{c,l,t-1})$. All control variables are described in Table 1.3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p $<$ 0.10, ** p $<$ 0.05, *** p $<$ 0.01.

Table B12: Robustness Check: Benchmark Regressions on a Balanced Panel

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.500**	0.563**	0.542***						
	(0.200)	(0.222)	(0.206)						
Cladistic similarity$_{t-1}$				0.491**	0.460*	0.407*			
				(0.231)	(0.238)	(0.238)			
Coethnic$_{t-1}$							0.328	0.337	0.338*
							(0.198)	(0.209)	(0.185)
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	177	177	177	177	177	177	177	177	177
Countries	23	23	23	23	23	23	23	23	23
Language groups	84	84	84	84	84	84	84	84	84
Adjusted R^2	0.921	0.921	0.921	0.921	0.920	0.921	0.920	0.920	0.921
Observations	3,894	3,894	3,894	3,894	3,894	3,894	3,894	3,894	3,894

This table reproduces benchmark estimates on a balanced subset of the panel dataset. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. All control variables are described in Table 1.3. Standard errors are clustered at the country-language group level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B13: Robustness Check: Benchmark Regressions Weighted by Language Group Population

Dependent Variable: $y_{c,l,t} = \ln(0.01 + \text{nightLights}_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.231**	0.329**	0.308**						
	(0.105)	(0.141)	(0.124)						
Cladistic similarity$_{t-1}$				0.202*	0.213**	0.190*			
				(0.103)	(0.108)	(0.107)			
Coethnic$_{t-1}$							0.161*	0.234**	0.260***
							(0.090)	(0.095)	(0.094)
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355	355	355	355
Countries	35	35	35	35	35	35	35	35	35
Language groups	163	163	163	163	163	163	163	163	163
Adjusted R^2	0.990	0.990	0.990	0.990	0.990	0.990	0.990	0.990	0.990
Observations	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610	6,610

This table reports the benchmark estimates weighted by Ethnologue language group population. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. All control variables are described in Table 1.3. Standard errors are clustered at the country-language group level and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B14: Robustness Check: Benchmark Regressions with Alternative Dependent Variables

	$\sqrt{\text{NightLights}_{c,l,t}}$			$\ln(NightLights_{c,l,t})$		
	(1)	(2)	(3)	(4)	(5)	(6)
Lexicostatistical similarity$_{t-1}$	0.038**			0.396**		
	(0.018)			(0.191)		
Cladistic similarity$_{t-1}$		0.029*			0.189	
		(0.016)			(0.163)	
Coethnic$_{t-1}$			0.012			0.258*
			(0.014)			(0.138)
Geographic controls	Yes	Yes	Yes	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	214	214	214
Countries	35	35	35	33	33	33
Language groups	164	164	164	98	98	98
Adjusted R^2	0.952	0.952	0.952	0.935	0.935	0.935
Observations	6,610	6,610	6,610	2,921	2,921	2,921

This table tests the robustness of the dependent variable using two alternative transformations: a square root of the raw night lights data ($\sqrt{\text{NightLights}_{c,l,t}}$) and the natural log of the raw night lights data without a constant term ($\ln(NightLights_{c,l,t})$). Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. All control variables are described in Table 1.3. Standard errors are clustered at the country-language group level and are reported in parentheses. * p $<$ 0.10, ** p $<$ 0.05, *** p $<$ 0.01.

Table B15: Individual-Level Regressions: Locational and Individual Similarity

Dependent Variable: DHS Wealth Index					
	(1)	(2)	(3)	(4)	(5)
Locational similarity$_{t-1}$	0.594	0.463***	0.479***	0.643***	0.365**
	(0.613)	(0.152)	(0.119)	(0.153)	(0.140)
Adjusted R^2	0.312	0.574	0.603	0.603	0.604
Individual similarity$_{t-1}$	1.260***	0.123	0.211	0.228	0.219
	(0.359)	(0.220)	(0.219)	(0.219)	(0.215)
Adjusted R^2	0.313	0.574	0.602	0.603	0.604
Locational similarity$_{t-1}$	0.592	0.463***	0.479***	0.643***	0.364**
	(0.613)	(0.153)	(0.119)	(0.153)	(0.140)
Individual similarity$_{t-1}$	1.259***	0.122	0.211	0.230	0.218
	(0.359)	(0.220)	(0.219)	(0.219)	(0.215)
Adjusted R^2	0.313	0.574	0.603	0.603	0.604
Rural indicator	No	Yes	Yes	Yes	Yes
Individual controls	No	No	Yes	Yes	Yes
Distance to border	No	No	No	Yes	No
Distance to coast	No	No	No	No	Yes
Country-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Locational language-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Individual language-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Clusters	88	88	88	88	88
Countries	13	13	13	13	13
Language groups	20	20	20	20	20
Observations	56,455	56,455	56,455	56,455	56,455

This table provides estimates for two channels: the effect of individual and locational similarity on the DHS wealth index. The unit of observation is an individual. The rural indicator is equal to 1 if a respondent lives in a rural location. The individual set of control variables include age, age squared, a gender indicator variable, an indicator for respondents living in the capital city, 5 education fixed effects and 7 religion fixed effects. Distance to the coast and border are in kilometers. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B16: Individual-Level Regressions: Locational and Individual Similarity

Dependent Variable: DHS Wealth Index					
	(1)	(2)	(3)	(4)	(5)
Locational coethnicity$_{t-1}$	0.838*	0.485***	0.437***	0.324**	0.601***
	(0.430)	(0.139)	(0.116)	(0.134)	(0.160)
Non-coethnic locational similarity$_{t-1}$	-0.692	0.348*	0.697***	0.573***	0.854***
	(0.556)	(0.205)	(0.148)	(0.167)	(0.173)
Rural indicator	No	Yes	Yes	Yes	Yes
Individual controls	No	No	Yes	Yes	Yes
Distance to coast	No	No	No	Yes	No
Distance to border	No	No	No	No	Yes
Country-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Locational language-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Individual language-wave fixed effects	Yes	Yes	Yes	Yes	Yes
Clusters	88	88	88	88	88
Countries	13	13	13	13	13
Language groups	20	20	20	20	20
Adjusted R^2	0.314	0.574	0.603	0.604	0.603
Observations	56,455	56,455	56,455	56,455	56,455

This table reports estimates that test for favoritism outside of coethnic language partitions. The unit of observation is an individual. The rural indicator is equal to 1 if a respondent lives in a rural location. The individual set of control variables include age, age squared, a gender indicator variable and an indicator for respondents living in the capital city. Distance to the coast and border are in kilometers. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B17: Individual-Level Regressions: Baseline Covariates

Dependent Variable: DHS Wealth Index									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Locational similarity$_{t-1}$	0.585	0.594	0.463***	0.636	0.490	1.024*	0.518	0.608	0.479***
	(0.604)	(0.613)	(0.152)	(0.398)	(0.637)	(0.592)	(0.587)	(0.399)	(0.119)
Age	-0.021***								-0.008
	(0.005)								(0.006)
Age squared	0.000***								0.000
	(0.000)								(0.000)
Female indicator		-0.010							0.112***
		(0.013)							(0.013)
Rural indicator			-1.846***						-1.606***
			(0.072)						(0.079)
Capital city indicator				1.502***					0.238***
				(0.053)					(0.053)
Distance to the coast					-0.001				
					(0.000)				
Distance to the border						-0.001*			
						(0.001)			
Religion FE	No	No	No	No	No	No	Yes	No	Yes
Education FE	No	No	No	No	No	No	No	Yes	Yes
Country-wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location-language-wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual-language-wave FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	88	88	88	88	88	88	88	88	88
Countries	13	13	13	13	13	13	13	13	13
Language groups	20	20	20	20	20	20	20	20	20
Adjusted R^2	0.316	0.312	0.574	0.342	0.314	0.317	0.317	0.416	0.603
Observations	56,455	56,455	56,455	56,455	56,455	56,455	56,455	56,455	56,455

This table establishes the impact of each baseline covariate used in Section 1.5. The unit of observation is an individual. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B18: Ethnic Favoritism and Coalition Power Sharing

	Share of cabinet positions			Share of top cabinet positions			Share of low cabinet positions		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Coethnicity$_t$	0.093***		0.100***	0.179***		0.185***	0.050***		0.057***
	(0.012)		(0.013)	(0.019)		(0.020)	(0.013)		(0.014)
Lexicostatistical similarity$_t$		0.095***			0.172***			0.057***	
		(0.013)			(0.021)			(0.013)	
Non-coethnic similarity$_t$			0.047**			0.047*			0.048**
			(0.018)			(0.022)			(0.019)
Group size controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Countries	15	15	15	15	15	15	15	15	15
Ethnic groups	187	187	187	187	187	187	187	187	187
Adjusted R^2	0.665	0.664	0.668	0.539	0.521	0.541	0.544	0.549	0.548
Observations	2,539	2,539	2,539	2,539	2,539	2,539	2,539	2,539	2,539

This table establishes that linguistic similarity predicts an ethnic group's share in the governing coalition of a country. The unit of observation is an ethnic group. The dependent variable in columns (1)-(3) is the share of cabinet positions of an ethnic group in the governing coalition, whereas in columns (4)-(6) and (7)-(9) the dependent variable measures the cabinet share of top positions and low positions. The group size controls include a time-invariant measure of an ethnic group's share of the national population and its polynomial. Standard errors are in parentheses and adjusted for clustering at the country level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix C

Chapter 2 Appendix

C.1 Variable Definitions, Summary Statistics and Data Sources

C.1.1 Data and Sources

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

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

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

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

Geographic distance: Geodesic distance between the most populated cities in a country pair,

measured per 100 kilometres. This data comes from a data set compiled by researchers at Centre d'Etudes Prospectives et d'Informations Internationales (CEPII).
Source: http://www.cepii.fr/cepii/en/welcome.asp

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

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

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

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

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

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

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

Human capital: The cross-country 5-year panel of education attainment for the total population aged 15 and over is from Barro and Lee (2013). Measures used include (i) the average years of

schooling attained, (ii) the percentage of complete primary schooling attained, (iii) the percentage of complete secondary schooling attained, and (iv) the percentage of complete tertiary schooling attained.
Source: http://www.barrolee.com

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

C.1.2 Summary Statistics

Table C1: Benchmark Sample Summary Statistics

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

Table C2: Commonly Translated Authors by Country

China	
Author	Subject
Leo Tolstoy	Literature
Dale Carnegie	Philosophy
Maxim Gorky	Literature

USA	
Author	Subject
Rudolf Steiner	Religion
Plato	Philosophy
Anton Chekhov	Literature

Brazil	
Author	Subject
Allan Kardec	Philosophy
Joseph Murphy	Religion
Agatha Christie	Literature

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

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

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

Argentina	
Author	Subject
José Trigueirinho Netto	Philosophy
Sigmund Freud	Psychology
Ramacharaka	Religion

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

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

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

Table C3: Language Pair Observations by Translating Country for the Benchmark Sample (1979-2005)

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

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

Table C4: Observations by Translating Language for the Benchmark Sample (1979-2005)

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

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

Table C5: Observations by Original Language for the Benchmark Sample (1979-2005)

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

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

C.2 Supplementary Materials

C.2.1 Alternative Measure of Linguistic Distance

In this section I reproduce the table of benchmark estimates using a cladistic measure of linguistic distance. I construct this measure of cladistic distance as outlined in Appendix A, which is measured as one minus the ratio of shared branches on the Ethnologue language tree. I assume the weighting scheme of Fearon (2003), where the ratio of shared tree branches is square rooted to discount more recent linguistic cleavages relative to deep cleavages. Because the number of branches varies among language familes and subfamilies, the maximum number of branches between any two languages is not constant. To overcome this obstacle I assume that all current languages are of equal distance from the proto-language at the root of the Ethnologue language tree. This assumption is equivalent to the assumption Desmet et al. (2012) use when constructing cladistic distances (see Figure A1).

Table C6 reports the benchmark estimates using this alternative measure. The two opposing forces of relatedness are borne out of this alternative data and estimated to be significantly different than zero. The effect of the cladistic measure of linguistic distance is smaller than the comparable lexicostatistical measure used in the benchmark estimate. This finding is consistent with the evidence in Dickens (2016a), where the added variation of the lexicostatistical measure yields added percision in estimation. Nonetheless, the basic finding of this paper is robust to alternative measures of linguistic distance.

C.2.2 Further Test of the Home Country Assumption

In this section I extend the analysis of section 2.4.1 and systematically dropping every original language of a translation with at least 100 observations in the benchmark sample. The idea here is to ensure that the benchmark results cannot be explained away by any one original language due to the ambiguity of the home country assignment rule. Table C7 reports these estimates and shows no evidence that any one language can explain away the benchmark result.

C.2.3 Dominant Book Subject?

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

To test this I re-estimate the preferred specification and drop all translations belonging to each subject area one at a time. The results are presented in Table C8. Coefficient estimates should

be interpreted relative to the benchmark estimate; i.e., small deviations from the benchmark estimate imply the excluded subject of translation is not influential over and above the average effect, whereas large deviations indicate subject areas that are particularly influential in the benchmark result.

The first observation about Table C8 is that no one subject area is driving the core result of this study – the standardized coefficient for linguistic distance range from -7.8 to -8.5 percent, and from 5.6 to 7.2 percent for genetic distance. It is reassuring that the benchmark estimate of -8.0 percent for linguistic distance and 6.3 percent for genetic distance tend towards the lower bound of this range of coefficients. Second, both genetic and linguistic distance are precisely estimated in all regressions at standard levels of confidence. These two observations suggest the baseline results of section 2.3.3 cannot be explained by a single outlier subject that is particularly responsive to either linguistic or genetic distance.

Table C6: Conditional Benchmark Regressions with Cladistic Linguistic Distance

	Dependent variable: Log translations					
	(1)	(2)	(3)	(4)	(5)	(6)
Linguistic distance (cladistic)	-1.77***	-1.76***	-1.76***	-1.41***	-1.41***	-1.31***
	(0.29)	(0.29)	(0.29)	(0.28)	(0.28)	(0.29)
Genetic distance	2.89***	2.87***	2.87***	3.05***	3.03***	2.16**
	(1.01)	(1.01)	(1.01)	(1.01)	(1.00)	(1.02)
Geographic distance	-0.89***	-0.89***	-0.89***	-0.60***	-0.57***	-1.35***
	(0.16)	(0.16)	(0.16)	(0.16)	(0.17)	(0.45)
Economic controls	No	Yes	Yes	Yes	Yes	Yes
Political controls	No	No	Yes	Yes	Yes	Yes
Trade controls	No	No	No	Yes	Yes	Yes
Colonial controls	No	No	No	No	Yes	Yes
Geography controls	No	No	No	No	No	Yes
Target Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Language FE	Yes	Yes	Yes	Yes	Yes	Yes
Target Country FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	39,275	39,275	39,275	39,275	39,275	39,275
Adjusted R^2	0.27	0.27	0.27	0.28	0.28	0.28
Country pair clusters	1897	1897	1897	1897	1897	1897
Standardized Coefficients (%)						
Linguistic Distance	-8.7	-8.7	-8.7	-7.0	-7.0	-6.5
Genetic Distance	7.6	7.5	7.5	8.0	7.9	5.7

This table re-produces the benchmark estimates of Table 2.3 using a cladistic measure of linguistic distance. The set of economic controls include log real GDP per capita and log population in both countries, political controls include political rights in both countries, trade controls include the logged value of bilateral trade flows of the country pair, the colonial controls include a dummy variable indicating if a country pair has ever been in a colonial relationship, and the geography controls include a set of indicators for contiguity and a country pair's absolute difference in latitude and longitude. Country-pair clustered robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C7: Sensitivity Analysis: Further Test of Home Country Assignment

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

This table tests the assignment of the home country of a language by systematically dropping each original language of a book translation with a 100 or more observations. The robustness of the results to this test suggests the benchmark result is not driven by the assumption of the original country of a translation. All regressions include the benchmark set of control variables used in column (6) of Table 2.3. All regressions also include individual country, year and target language fixed effects. Country-pair clustered robust standard errors reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C8: Robustness Check for Dominant Subject of Translation

	Dependent variable: Log translations							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Excluding:	Arts	History	Literature	Philosophy	Religion	Social Science	Applied Science	Natural Science
Linguistic distance	-1.09***	-1.13***	-1.12***	-1.11***	-1.08***	-1.18***	-1.13***	-1.11***
	(0.28)	(0.28)	(0.34)	(0.28)	(0.28)	(0.28)	(0.27)	(0.28)
Genetic distance	2.33**	2.48**	2.47*	2.13**	2.39**	2.78***	2.43**	2.47**
	(1.07)	(1.03)	(1.29)	(1.04)	(1.09)	(1.03)	(1.02)	(1.06)
Geographic distance	-1.34***	-1.32***	-1.54***	-1.42***	-1.36***	-1.44***	-1.33***	-1.33***
	(0.45)	(0.46)	(0.53)	(0.45)	(0.46)	(0.45)	(0.45)	(0.45)
Benchmark controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Target Language FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original Language FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Target Country FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	38,405	37,661	26,488	38,610	35,471	37,717	38,374	38,987
Adjusted R^2	0.29	0.30	0.31	0.29	0.30	0.30	0.29	0.28
Country pair clusters	1887	1873	1505	1885	1805	1852	1884	1893
Standardized Coefficients (%)								
Linguistic Distance	-7.8	-8.1	-8.2	-8.0	-7.8	-8.5	-8.2	-8.0
Genetic Distance	6.1	6.5	6.6	5.6	6.2	7.2	6.4	6.5

This table reports estimates on various subsamples that exclude each subject classification in the data one by one. All regressions include the benchmark set of control variables used in column (6) of Table 2.3. Country-pair clustered robust standard errors reported in parentheses. * p $<$ 0.10, ** p $<$ 0.05, *** p $<$ 0.01.

Table C9: Ethnologue Home Country vs. Synthetic Language Country Assignment

	Ethnologue Home Country Mean	Synthetic Language Country Mean	Difference	Observations
Genetic distance	0.041	0.045	-0.004***	35,732
	(0.000)	(0.000)	(0.000)	
Geographic distance	0.390	0.455	-0.066***	35,737
	(0.002)	(0.002)	(0.001)	
Log real GDP	12.989	13.222	-0.233***	37,976
	(0.007)	(0.008)	(0.003)	
Log population	3.457	4.133	-0.675***	37,976
	(0.008)	(0.009)	(0.007)	
Political rights	0.163	0.171	-0.008***	37,976
	(0.001)	(0.001)	(0.000)	
Log bilateral trade	5.932	5.972	-0.041***	37,549
	(0.012)	(0.012)	(0.003)	
Colonial relationship	0.093	0.075	0.017***	35,737
	(0.002)	(0.001)	(0.001)	
Contiguity	0.116	0.112	0.004***	35,737
	(0.002)	(0.002)	(0.001)	
Abs. difference in latitude	14.751	16.142	-1.391***	37,976
	(0.091)	(0.094)	(0.054)	
Abs. difference in longitude	40.893	46.247	-5.354***	37,976
	(0.226)	(0.238)	(0.095)	

This table establishes the mean difference in covariates between the alternate home country assignment rules of a book translation.

Appendix D

Chapter 3 Appendix

D.1 Data Description and Sources

Ethnolinguistic groups: Georeferenced group data comes from the World Language Mapping System (WLMS). 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/

Caloric suitability index: I sourced the caloric suitability index (CSI) data from Galor and Özak (2016). CSI is a measure of agricultural productivity that reflects the caloric potential in 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 CSI measure that includes cells with zero productivity. I measure average CSI within each buffer zone using Goode's homolosine map projection to minimize area distortions. I use Stata to calculate the the post-1500 change in CSI .
Source: http://omerozak.com/csi

Agricultural suitability index: I sourced the agricultural suitability index (ASI) data from Ramankutty et al. (2002). ASI is an index of the suitability of land for agriculture, which measures the fraction of each 0.5×0.5 grid cell that is suitable for agriculture. I measure average ASI within each buffer zone using Goode's homolosine map projection to minimize area distortions.
Source: https://goo.gl/pPNxLVi

Absolute difference in elevation: I collect elevation data from the National Geophysical Data Centre (NGDC) at the National Oceanic and Atmospheric Administration (NOAA). I measure average elevation of each buffer zone using Goode's homolosine map projection.
Source: www.ngdc.noaa.gov/mgg/topo/globe.html

Ruggedness: As a measure of ruggedness I use the standard deviation of the NGDC elevation data. I use Goode's homolosine map projection.
Source: www.ngdc.noaa.gov/mgg/topo/globe.html

Precipitation: Precipitation data comes from the WorldClim – Global Climate Database. I measure average precipitation within each buffer zone using Goode's homolosine map projection to minimize area distortions.
Source: http://www.worldclim.org/current

Temperature: Temperature data comes from the WorldClim – Global Climate Database. I measure the average temperature within each buffer zone using Goode's homolosine map projection to minimize area distortions.
Source: http://www.worldclim.org/current

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 construct a measure of the average prevalence of malaria with each buffer zone using Goode's homolosine map projection.
Source: https://sites.google.com/site/gordoncmccord//datasets

Lakes and Rivers: Georeferenced data on lakes and rivers from around the world come from Natural Earth. I use the Equidistant Cylindrical projection in ArcGIS, and identify buffer zones that intersect with a river or lake. I use Stata to construct two indicator variables: one indicator when a lake intersect the buffer zone and another when a river intersects the buffer zone.
Source: http://www.naturalearthdata.com/downloads/10m-physical-vectors/

Area of language pair: Total area of neighbouring ethnolinguistic group polygons, measured in kilometres squared.
Source: Calculated using ArcGIS.

Population of language pair: Ethnolinguistic group population comes from the WLMS Ethnologue database. Population of language pair is the sum of both groups population.
Source: Calculated using Stata.

Latitude and Longitude difference: Latitude and longitude coordinates for an ethnolinguistic group correspond to the group's centroid. Differences are calculated by taking the absolute difference of a neighbouring pair centroids.
Source: Calculated using ArcGIS and Stata.

D.2 Supplementary Material

Table D1 reports summary statistics for the full sample dataset and Table D2 reports summary statistics for the sibling sample dataset. Table D3 is a replication of Table 3.4, but includes the coefficient estimates for all control variables. Similarly, Table D4 is a replication of Table 3.5 that includes all coefficient estimates.

Table D1: Summary Statistics – Full Sample

	Mean	Std dev.	Min	Max	N
Linguistic distance	0.727	0.180	0.003	0.945	6,990
CSI variation (pre-1500)	0.222	0.265	0.000	1.616	6,990
Change in CSI variation (post-1500)	-0.063	0.181	-0.977	0.321	6,990
CSI (pre-1500)	1.346	0.706	0.000	4.939	6,990
Change in CSI (post-1500)	-0.061	0.572	-2.136	1.361	6,990
Malaria	8.238	8.923	0.000	36.29	6,990
Ruggedness	0.299	0.292	0.000	1.934	6,990
Elevation	0.696	0.662	-0.022	4.929	6,990
Precipitation	13.65	7.964	0.000	50.67	6,990
Precipitation variation	1.683	1.846	0.000	19.65	6,990
Temperature	21.54	6.917	-12.29	29.49	6,990
Temperature variation	1.552	1.502	0.000	10.09	6,990
River	0.612	0.487	0.000	1.000	6,990
Lake	0.096	0.295	0.000	1.000	6,990
Area of language pair (km^2)	10.05	2.384	5.353	16.00	6,990
Population of language pair	13.00	3.18	5.298	20.64	6,990
Latitude difference	1.505	2.515	0.000	30.34	6,990
Longitude difference	2.151	7.201	0.000	340.3	6,990
Land suitability variation	0.091	0.077	0.000	0.469	6,966

Table D2: Summary Statistics – Sibling Sample

	Mean	Std dev.	Min	Max	N
Linguistic distance	0.514	0.175	0.003	0.913	1,277
CSI variation (pre-1500)	0.266	0.298	0.000	1.373	1,277
Change in CSI variation (post-1500)	-0.11	0.239	-0.977	0.220	1,277
CSI (pre-1500)	1.383	0.702	0.000	4.598	1,277
Change in CSI (post-1500)	-0.119	0.643	-2.064	1.335	1,277
Malaria	8.822	8.847	0.000	35.62	1,277
Ruggedness	0.337	0.311	0.000	1.897	1,277
Elevation	0.716	0.662	0.000	4.929	1,277
Precipitation	15.24	8.306	0.000	48.68	1,277
Precipitation variation	1.879	1.841	0.000	16.69	1,277
Temperature	22.17	5.902	-6.977	29.21	1,277
Temperature variation	1.734	1.596	0.000	9.981	1,277
River	0.524	0.500	0.000	1.000	1,277
Lake	0.081	0.272	0.000	1.000	1,277
Area of language pair (km^2)	8.694	1.892	5.376	15.97	1,277
Population of language pair	11.33	2.664	5.298	20.64	1,277
Latitude difference	0.642	1.12	0.000	15.23	1,277
Longitude difference	0.734	2.261	0.000	51.60	1,277
Land suitability variation	0.082	0.073	0.000	0.434	1,267

Table D3: Robustness Check: Native Population Sensitivity Analysis

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$						
	& >25% Native Pop.		& >50% Native Pop.		& >75% Native Pop.	
	(1)	(2)	(3)	(4)	(5)	(6)
CSI Variation (pre-1500)	-0.121***	-0.200***	-0.121***	-0.212***	-0.132***	-0.252***
	(0.037)	(0.065)	(0.037)	(0.065)	(0.039)	(0.070)
Change in CSI Variation (post-1500)	-0.209***	-0.248***	-0.218***	-0.269***	-0.202***	-0.281***
	(0.055)	(0.070)	(0.055)	(0.070)	(0.057)	(0.074)
CSI (pre-1500)	-0.025	-0.023	-0.023	-0.014	-0.008	-0.008
	(0.015)	(0.016)	(0.015)	(0.016)	(0.016)	(0.017)
Change in CSI (post-1500)	-0.001	-0.012	-0.004	-0.016	0.001	-0.008
	(0.024)	(0.026)	(0.023)	(0.026)	(0.024)	(0.027)
Malaria		0.001		0.002		0.002
		(0.001)		(0.001)		(0.001)
Ruggedness		0.195**		0.188*		0.185
		(0.096)		(0.097)		(0.128)
Elevation		0.023		0.025		0.030
		(0.019)		(0.019)		(0.021)
Precipitation		-0.001		-0.001		-0.001
		(0.001)		(0.001)		(0.002)
Precipitation Variation		0.003		0.004		0.008
		(0.005)		(0.005)		(0.006)
Temperature		0.003**		0.003**		0.003
		(0.002)		(0.002)		(0.002)
Temperature Variation		-0.028		-0.026		-0.027
		(0.018)		(0.018)		(0.023)
River		-0.036***		-0.035***		-0.045***
		(0.011)		(0.011)		(0.013)
Lake		-0.054**		-0.056**		-0.057**
		(0.022)		(0.022)		(0.027)
Area of Language Pair (km^2)		-0.009		-0.009		-0.005
		(0.006)		(0.006)		(0.007)
Population of Language Pair		-0.004		-0.003		-0.005
		(0.004)		(0.004)		(0.005)
Latitude Difference		0.007		0.007		0.005
		(0.006)		(0.006)		(0.006)
Longitude Difference		0.005		0.004		0.002
		(0.004)		(0.003)		(0.003)
Latitude × Longitude		-0.000		0.000		0.000
		(0.000)		(0.000)		(0.000)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Continental FE	Yes	Yes	Yes	Yes	Yes	Yes
Language Cluster 1	915	915	886	886	767	767
Language Cluster 2	826	826	795	795	693	693
Observations	1,241	1,241	1,202	1,202	1,036	1,036
Adjusted R^2	0.34	0.39	0.34	0.39	0.33	0.37

This table tests the sensitivity of the benchmark estimates by limiting the sibling sample to pairs that reside in a country where a significant portion of their population is native to that country. Columns (1)-(2), (3)-(4) and (5)-(6) report estimates where at least 25, 50 and 75 percent of the population is native to the country of residence, respectively. The unit of observation is a 100 km buffer zone along the contiguous border segment of each language pair. 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 D4: Robustness Check: Overlapping and Multi-Part Polygon Sensitivity Analysis

Dependent Variable: Lexicostatistical Linguistic Distance $\in (0, 1)$						
	Excluding Overlapping Groups		Excluding Multi-Part Groups		Excluding Overlapping & Multi-Part Groups	
	(1)	(2)	(3)	(4)	(5)	(6)
CSI Variation (pre-1500)	-0.100**	-0.197***	-0.117***	-0.215**	-0.136***	-0.205*
	(0.041)	(0.073)	(0.045)	(0.086)	(0.050)	(0.106)
Change in CSI Variation (post-1500)	-0.175***	-0.221***	-0.218***	-0.262***	-0.213***	-0.244**
	(0.061)	(0.080)	(0.071)	(0.099)	(0.076)	(0.118)
CSI (pre-1500)	-0.022	-0.023	-0.034	-0.036*	-0.024	-0.027
	(0.019)	(0.020)	(0.021)	(0.022)	(0.027)	(0.030)
Change in CSI (post-1500)	-0.000	-0.018	-0.011	-0.019	-0.009	-0.019
	(0.027)	(0.030)	(0.033)	(0.036)	(0.038)	(0.041)
Malaria		0.002		0.001		0.002
		(0.001)		(0.002)		(0.002)
Ruggedness		0.184*		0.163		0.150
		(0.108)		(0.116)		(0.123)
Elevation		0.020		0.039		0.041
		(0.024)		(0.024)		(0.042)
Precipitation		-0.003*		-0.001		-0.002
		(0.002)		(0.002)		(0.002)
Precipitation Variation		0.006		0.000		0.001
		(0.006)		(0.007)		(0.008)
Temperature		0.003		0.007**		0.008
		(0.002)		(0.003)		(0.006)
Temperature Variation		-0.023		-0.015		-0.015
		(0.020)		(0.022)		(0.024)
River		-0.032**		-0.014		-0.013
		(0.013)		(0.015)		(0.017)
Lake		-0.041		-0.027		-0.023
		(0.025)		(0.030)		(0.036)
Area of Language Pair (km^2)		-0.003		-0.003		-0.005
		(0.007)		(0.010)		(0.012)
Population of Language Pair		-0.008		-0.009		-0.004
		(0.005)		(0.007)		(0.008)
Latitude Difference		-0.003		-0.027		-0.014
		(0.008)		(0.023)		(0.033)
Longitude Difference		0.003		-0.021		-0.027
		(0.004)		(0.023)		(0.036)
Latitude × Longitude		0.000		0.012		0.015
		(0.000)		(0.010)		(0.023)
Language Family FE	Yes	Yes	Yes	Yes	Yes	Yes
Continental FE	Yes	Yes	Yes	Yes	Yes	Yes
Language Cluster 1	744	744	587	587	492	492
Language Cluster 2	666	666	539	539	453	453
Observations	988	988	763	763	632	632
Adjusted R^2	0.36	0.40	0.36	0.40	0.37	0.39

This table tests the sensitivity of the benchmark estimates by limiting the sibling sample to ethnolinguistic pairs that do not overlap with any other groups and are not composed of multi-part group polygons. The unit of observation is a 100 km buffer zone along the contiguous border segment of each language pair. Standard errors are double-clustered at the level of each language group and are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.