Ethnolinguistic Favoritism in African Politics

Andrew Dickens[†]

10 August 2016

I document evidence of ethnic favoritism in 164 language groups across 35 African countries using a new computerized lexicostatistical measure of relative similarity between each language group and their incumbent national leader. I measure patronage with night light luminosity, and estimate a positive effect of linguistic similarity off of changes in the ethnolinguistic identity of a leader. Identification of this effect comes from exogenous within-group time-variation among language groups partitioned across national borders. I then corroborate this evidence using survey data and establish that the benefits of favoritism result from a region's associated ethnolinguistic identity and not that of the individual respondent.

[†]York University, Department of Economics, Toronto, ON. E-mail: adickens@yorku.ca.
I am indebted to Nippe Lagerlöf for his encouragement and detailed feedback throughout this project. I thank Matthew Gentzkow and two anonymous referees for helpful suggestions that have greatly improved this paper. I also thank Tasso Adamopoulos, Greg Casey, Mario Carillo, Berta Esteve-Volart, Raphaël Franck, Oded Galor, Fernando Leibovici, Stelios Michalopoulos, Stein Monteiro, Laura Salisbury, Ben Sand, Assaf Sarid and David Weil for helpful comments, in addition to seminar participants at the Brown University Macro Lunch, the Royal Economic Society's 2^{nd} Symposium for Junior Researches, the PODER Summer School on "New Data in Development Economics", the Canadian Economics Association Annual Conference and York University. This research is funded by the Social Science and Humanities Research Council of Canada. All errors are my own.

1 Introduction

Ethnolinguistic group affiliation is a salient marker of identity in Africa. People identify as coethnics because they share a common ancestry and language, hold similar cultural beliefs and pursue related economic activities. Yet these shared characteristics are as much a badge of group identity as they are a means to discrimination. Ethnic group divisions are especially problematic in regions like Africa, where high levels of diversity constrain the economic performance and political functioning of a country (Easterly and Levine, 1997; Alesina et al., 2003; Alesina and La Ferrara, 2005; Ashraf and Galor, 2013). Demarcation by ethnic identity is particularly evident in African politics, where group affiliation is an important factor in how political favors are allocated (Franck and Rainer, 2012; Kramon and Posner, 2014; Burgess et al., 2015).

In this paper I revisit the study of ethnic favoritism in Africa with a new measure of group relatedness – linguistic similarity. Language is perhaps the most visible marker of ethnic identity in Africa, and to its advantage language can be observed as a continuous measure of relatedness. The conventional approach is to study coethnicity, a binary outcome that cannot account for the relative proximity of a leader to all groups that do not share an identical ethnolinguistic background. Language is also advantageous because it does not evolve in isolation from biological and social factors, and so linguistic similarity should be interpreted as an implicit measure of a whole set of ancestral and cultural traits that are important to group identity.[1]

The introduction of this new measure also provides testable grounds for the central hypothesis of this paper: a group that is linguistically similar to the ethnolinguistic identity of their national leader will be better off than a more distant group. I provide robust empirical evidence of this effect in 164 subnational language groups across 35 African countries, a phenomenon I term ethnolinguistic favoritism.[2] I also find that relative differences in language

[1]The co-evolution of language and genetics has been of recent interest to economists (Spolaore and Wacziarg, 2013, 2015) and has a long history in population genetics (Boyd and Richerson, 1985).

[2]I use the term ethnolinguistic favoritism because there is a close mapping between

matter too, which suggests that leaders exhibit a continuum of group preference that is increasing in similarity. This latter finding builds on existing evidence that favoritism is a binary phenomenon.

Measuring favoritism in African language groups is inherently difficult because these groups are routinely small and do not correspond to subnational administrative regions. To overcome this challenge I use satellite imagery of night light luminosity as an aggregate measure of patronage in African language groups between 1992–2013.[3] This broad measure of patronage is advantageous because I study a large sample of countries. Kramon and Posner (2013) examine multiple patronage goods in six African countries and find no evidence among these goods of a consistent pattern of patronage. Hence my use of night light intensity has the advantage that it overcomes this problem of external validity inherent to multi-country studies of a single patronage good.

To determine the ethnolinguistic identity of a leader I use a variety of sources. To start I map the spatial distribution of African languages groups and use geo-referenced birthplace coordinates to identify the language group associated with a political leader's birthplace, an approach similar to Hodler and Raschky (2014). Then I collect information on the ethnic identity of a leader using data from Dreher et al. (2014), Francois et al. (2015) and Wikipedia. When the birth language and ethnic identity share an identical name I assign the corresponding ethnolinguistic identity. When they differ I check if the geo-referenced birth language is a language of the identified ethnic group; if so I assign the language as the leader's ethnolinguistic identity, and if there is no clear association between the birth language and ethnicity, I assign the dominant language of the leader's ethnic group found in the country.

For each year of the panel I use a computerized lexicostatistical estimate of phonological relatedness to measure the similarity of each language group to their national leader. This new measure extends the conventional measure of coethnicity with the added precision of measured similarity between each

language and ethnicity in Africa (Batibo, 2005; Desmet et al., 2015).

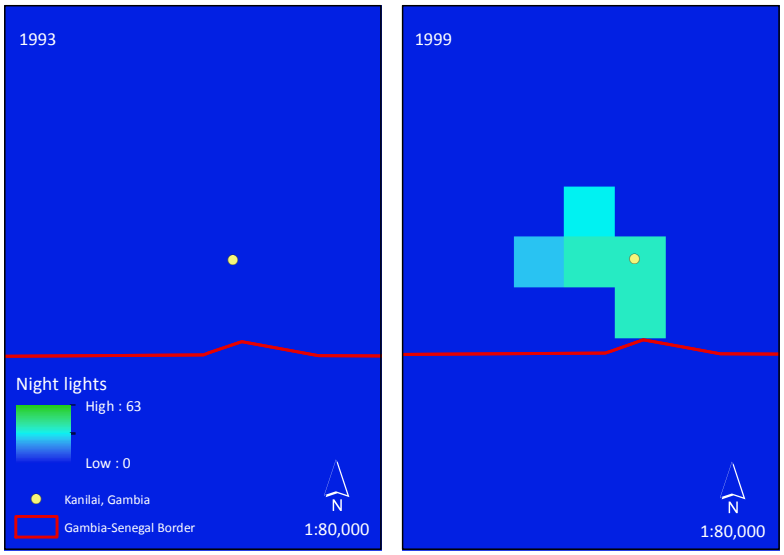
[3]Henderson et al. (2012) document a strong relationship between night light intensity and GDP at the national level, and Hodler and Raschky (2014) show this to be true as the subnational level as well.

leader and non-coethnic language group. To identify the effect of linguistic similarity I restrict the analysis to partitioned language groups. I define a partition as a language group separated by one or more national borders. The historical formation of these partitions began with the Berlin Conference of 1884-1885, where European powers divided up Africa with little regard for the homelands of these ethnolinguistic groups (Herbst, 2000). This disregard led to the arbitrary formation of national borders, which ironically "did not reflect reality but helped create it" (Wesseling, 1996, p.364). One such reality was the partitioning of nearly 200 language groups throughout Africa.

In the context of this study, the quasi-random nature of African border design identifies exogenous variation in similarity because the ethnolinguistic identity of a national leader varies across borders within the same partitioned language group. Because a language group shares a common ancestry, and is 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 fraction of that same group on the other side of the border. The use of African borders as a source of exogenous variation is methodologically similar to Michalopoulos and Papaioannou (2014).

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 the next several years little changed in Senegal as President Diouf's reign continued. On the contrary, much changed for the Jola-Fonyi of Gambia 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 has 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: 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.

Figure 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 Gambia 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 is exactly the within-group variation that I identify off of in my benchmark estimates. 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.

In my benchmark specification, I estimate an average effect of similarity off of changes in the ethnolinguistic identity of a leader using a triple difference-in-differences estimator. I find that a standard deviation increase in linguistic similarity (23 percent change) yields a 3 percent increase in group-level GDP. 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 exploit the continuity of lexicostatistical similarity by separating out the binary effect of coethnicity from non-coethnic similarity, and find evidence of favoritism in these non-coethnic regions. This evidence supports my hypothesis of favoritism working across a gradient of group similarity in Africa.

To address the robustness of the benchmark estimates I run a series of tests. I directly control for the geodesic distance and the difference in geographic endowments of each group to the leader's ethnolinguistic homeland (Michalopoulos, 2012). I also show that the benchmark estimate is robust to controlling for lagged values of night light intensity, which suggests that the economic success of a leader's ethnolinguistic group prior to election does not drive the results. To the contrary there is no statistical relationship between linguistic similarity and night light activity when analyzing the similarity of a language group to their leader in the period immediately before a leadership change. This lack of a pre-trend in the data is reassuring that the common trends assumption of my identification strategy is satisfied.

Next I examine the dynamics of incumbency. I find that my benchmark result is largely driven by leaders who've held office longer than the sample median of nine years. The policy implication of this finding is that incumbency term limits can be used as a tool to minimize the extent of favoritism throughout Africa.

After I establish the robustness of the benchmark estimates, I study two channels through which leaders may plausibly provide favors. I distinguish between the conventional mechanism that links a leader's investment in a region to that region's ethnolinguistic identity, and an individual-level mechanism where linguistic similarity affords an individual increased opportunity irrespective of where they live. To do this I use individual-level data from the Demographic and Health Survey (DHS) and construct two measures of lexicostatistical similarity: the similarity of a leader's ethnolinguistic identity to the language region in which a respondent resides, and the similarity of that respondent's home language to the incumbent leader. The inclusion of both measures in an estimating equation exploits the fact that individuals who are non-native to the region they inhabit will vary in terms of their individual and regional linguistic similarity.

To replicate the same source of variation used in the regional-level analysis, I project the DHS cluster coordinates onto a map of partitioned language groups to pinpoint respondents living in a partitioned language group. I identify partitioned language groups with DHS clusters available on both sides of the border, and use those with at least 2 consecutive surveys from the same set of DHS waves. Doing so allows me to exploit time-variation across waves within a partitioned language group at the individual level. With this set-up I find strong evidence of the locational channel, but only weak evidence of the individual channel. This indicates that leaders provide favors through regional transfers, and that these transfers are beneficial to all inhabits of a region regardless of their background. Again I show that this is not entirely a coethnic effect, and that relative similarity in language matters too.

These results contribute to a body of evidence that favoritism works across ethnolinguistic lines in Africa (Golden and Min, 2013). My first contribution is evidence of favoritism working across a gradient of similarity. The innovation of a lexicostatistical measure of similarity is the observable variation in similarity between a leader and non-coethnic groups that is not possible using a conventional binary measure of coethnicity. The superiority of this new measure is evidenced by the added precision it yields in estimation. While this

approach is new to the ethnic favoritism literature, this computerized lexicostatistical measure has been elsewhere.[4] A number of other economists have used data from Dyen et al. (1992) to measure lexicostatistical linguistic distances (Desmet et al., 2005, 2009, 2011; Spolaore and Wacziarg, 2009, 2014, 2016), but this data is restricted to 84 Indo-European language distances only, none of which represent a historical language group of Africa. The Dyen et al. (1992) data also does not employ the computerized estimation approach used here, and instead relies on subjective judgements of similarity.[5]

My second contribution is methodological. To the best of my knowledge, this is the first empirical study of ethnic favoritism in Africa to exploit *within*-group differences as a way of accounting for the complete history of a group. The novelty of this within-group variation is that I can use one fraction of a partitioned language group as a counterfactual observation for the fraction of that *same* group on the other side of the border. Doing so accounts for the long-run persistence of a group's pre-colonial history on that group's political and economic success today (Gennaioli and Rainer, 2007; Michalopoulos and Papaioannou, 2013; Fenske, 2013). Michalopoulos and Papaioannou (2014) similarly use the colonial partitioning of Africa as a source of variation, but instead study the role of national institutions on subnational development. Using a cross-section of ethnic groups they identify off of variation in the level of institutions within a partitioned group. To the contrary I possess time variation at the group level, which allows me to identify off of *changes* in the ethnolinguistic identity of a leader while flexibly controlling for all country-group-level features.

Closer to my paper in content is Franck and Rainer's (2012) study of eth-

[4]For example, this lexicostatistical measure 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, 2016). See Ginsburgh and Weber (2016) for a discussion of this linguistic distance measure.

[5]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 (i.e., having a shared origin). This approach relies on subjectively determined cognates, and is extremely labor intensive, thus inhibiting the number of language comparisons possible. Greenberg (1956) formally introduced this approach, and Dyen et al. (1992) discuss the procedure.

nic favoritism in Africa. These authors study the benefits of coethnicity with between-group comparisons within a country, where different ethnic groups serve as the counterfactual observation. What sets my paper apart from Franck and Rainer's (2012) is that all cultural and geographic features of an ethnolinguistic group are held constant across counterfactual observations in my empirical model. The ubiquity of group partitions in Africa also enables me to expand Franck and Rainer's (2012) evidence of favoritism in 18 African countries to 35 countries. More commonly researchers focus on a single patronage good in a single country. For example, Kramon and Posner (2014) 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. In a recent manuscript, De Luca et al. (2015) extend the analysis beyond Africa with the proposition that ethnic favoritism is an axiom of politics and not simply an African phenomenon.[6]

My third contribution is the separate estimation of two channels of causality: a locational channel and a preferential access channel. Using survey data I separate the effect of locational similarity from individual similarity using variation among non-native individuals who live outside of their ethnolinguistic homeland. I examine these two mechanisms to test whether individuals are better off because the region in which they reside is historically similar to the ethnolinguistic background of their leader, or because they themselves are similar to their leader and benefit from favoritism irrespective of where they live. I find strong evidence of the locational channel, but only weak evidence of the individual-level channel.

The rest of this paper is structured as follows. Section 2 describes how I identify language group partitions and measure linguistic similarity. This section also previews the empirical results with patterns in the data. Section 3

[6]Yet consensus on African favoritism has not been reached. Francois et al. (2015) provide theoretical and empirical evidence that leader's provide only a small premium to their coethnics, and otherwise political power is proportional to group size in Africa. Kasara (2007) finds that African leaders extract more tax money from their own ethnic homelands because they have a better understanding of internal markets in that region.

Figure 2: Language Groups

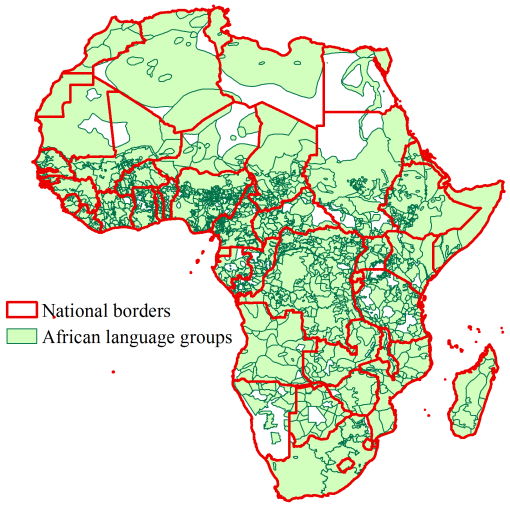
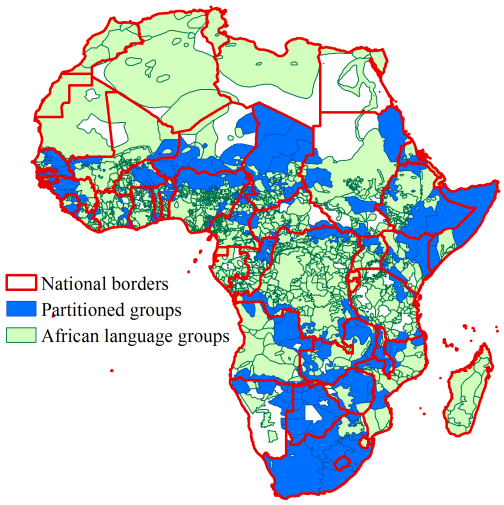


Figure 3: Language Partitions



outlines the empirical model and identification strategy, and Section 4 reports the benchmark estimates and robustness checks. In Section 5 I contrast the effects of locational and individual similarity using survey data. Section 6 concludes.

2 Data

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

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 2). In total there are 2,391 country-language group observations reflecting 1,956 unique language groups in 47 continental African countries.[7]

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

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. I drop all country-language groups whose reported Ethnologue population is zero since this would imply an absence of the group's presence in their homeland. The result is 486 remaining country-language group observations, made up of 227 unique language groups partitioned across 37 African countries.

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. These data have proliferated in economics research because of two features: night lights data exhibit a strong empirical relationship with GDP (Henderson et al., 2012), and because these data are available at a spatial resolution of 30-arc seconds (approximately 1 square kilometre).[8] The fine resolution of these lights data facilitates a proxy measure of GDP 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 measure of economic activity 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 only process observations that fall within the dark half of the lunar cycle, the period of time when the sun sets early. Scientists also remove any light activity related to the northern and southern lights, forest fires and imagery affected by cloud cover. All daily images that pass this screening process are then averaged for the entire year producing a

[8] 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.

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.[9]

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.

2.3 Assignment of a Leader's Ethnolinguistic Identity

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 propose a three-step assignment rule to construct a mapping between ethnicity and language in Africa. I require two key pieces of information: the language group associated with a leader's birthplace and each leader's ethnic identity.

I locate the birthplace of a leader using Wikipedia, and collect latitude and longitude coordinates for each birthplace from www.latlong.net. I then project these coordinates onto the Ethnologue map of Africa to back out the language group associated with each leader's birthplace.[10] I exclude leader's born abroad (4 leaders) since their ethnolinguistic group is not home to the country they govern.[11] To identify a leader's ethnic identity I use data from Dreher et al. (2014) 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 Wikipedia.

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

[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).

Using these data I adhere to the following three-step assignment rule to assign a leader's ethnolinguistic identity.

Step 1: I compare the birthplace linguistic identity with the ethnic identity for the 97 leaders in my benchmark sample. In 51.5 percent of the sample the birth language and ethnic identity are equivalent (50 leaders). For these leaders I assign the birthplace language as the leader's ethnolinguistic identity.

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 28.9 percent of the sample this is true (28 leaders); once again I assign the birthplace language as the leader's ethnolinguistic identity.

Step 3: For the remaining sample of unmatched leaders, I search the list of languages in each leader's country and match a language if its name is identical to the leader's ethnicity. In 14.4 percent of the sample this is true (14 leaders). For the remaining 5.2 percent of leaders (5 leaders), I assign the most populated language group associated with a leader's ethnic group.

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

[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.

easily computed for any language pair, since language trees exist for virtually all known world language families (Lewis, 2009).[13]

In this paper I use a computerized lexicostatistical measure of linguistic similarity developed by the Automatic Similarity Judgement Program (ASJP). 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]

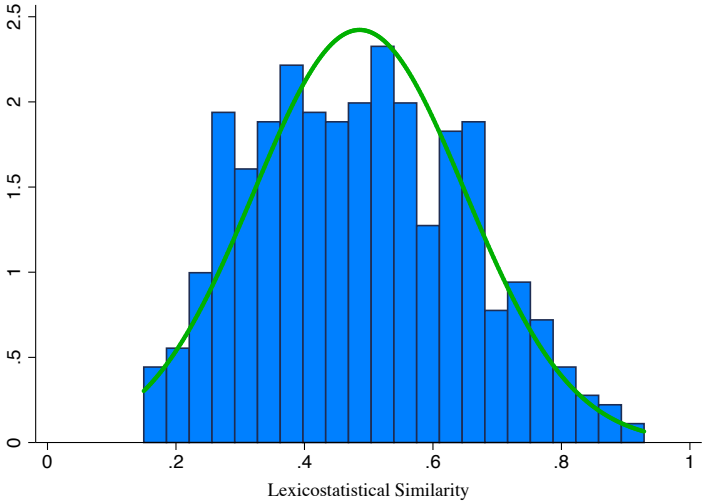
Then for each language pair of interest I run the Levenshtein distance algorithm on the respective language lists, which calculates the minimum number of edits necessary to translate the spelling of each word from one language to another. To correct for the fact that longer words will demand more edits, 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

[13] See Appendix B for a more detailed discussion of how cladistic similarity is measured.

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

Figure 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.

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 B.

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 those language pairs that share a common parent language on the Ethnologue language tree. Let these language pairs be known as siblings. In all cases these sibling pairs share the maximum number of tree nodes and have no observable difference in cladistic similarity

among them. To the contrary these sibling pairs exhibit substantial variation in lexicostatistical similarity. To make this point clear, I plot the distribution of lexicostatistical similarities among African sibling language pairs in Figure 4. This figure highlights the sizeable dispersion in lexicostatistical similarities even among sibling language pairs.

2.4.1 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 164 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. The result is an (unbalanced) panel of lexicostatistical similarity between partitioned language groups and their national leader for the years 1992-2013.[15] Figure 3 colour codes these partitioned language groups in blue. T he 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.

2.5 Patterns in the Data

Table 1 reports descriptive statistics for the night lights and language data. For completeness, I've included a cladistic measure of similarity and a binary measure of coethnicity.[16] 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

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

[16] 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.

Table 1: Descriptive Statistics

	Obs.	Mean	Std. Dev.	Min	Max
ln(0.01 + night lights)	6,475	-3.496	1.423	-4.605	1.515
Lexicostatistical similarity	6,475	0.192	0.229	0.000	1.000
Cladistic similarity	6,475	0.411	0.327	0.000	1.000
Coethnicity	6,475	0.046	0.210	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 the Data Appendix A for a description of the data and sources.

Table 2: Means of Linguistic Similarity Above-Below Median Night Lights

	Above Median Luminosity	Below Median Luminosity	Difference	p-value
Panel A: Full Sample				
Lexicostatistical similarity	0.247	0.137	0.109	0.000
	(0.005)	(0.003)	(0.006)	
Cladistic similarity	0.483	0.338	0.145	0.000
	(0.006)	(0.005)	(0.008)	
Coethnicity	0.080	0.012	0.067	0.000
	(0.005)	(0.002)	(0.005)	
Panel B: Non-Coethnic Regions Only				
Lexicostatistical similarity	0.181	0.125	0.057	0.000
	(0.003)	(0.002)	(0.004)	
Cladistic similarity	0.439	0.325	0.114	0.000
	(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 6,475 observations and the non-coethnic subsample consists of 6,177 observations. Standard errors are reported in parentheses.

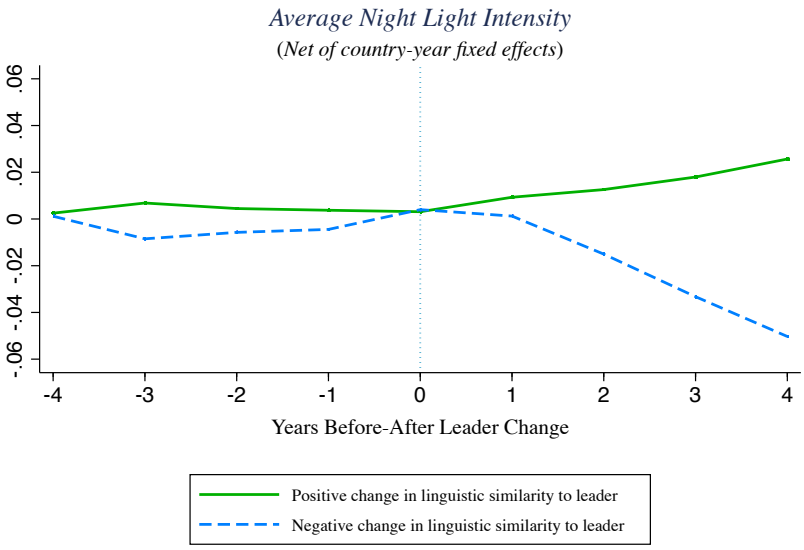
benchmark sample. Doing so helps correct for the non-normal nature of the data and preserves sample size, and allows for a (near) semi-elasticity interpretation of the benchmark empirical model. The mean value of lexicostatistical linguistic similarity says that country-language groups are 19.2 percent similar to their national leader on average, and the mean value of cladistic similarity implies 41.1 percent similarity. The mean value of coethnicity says that 4.6 percent of the benchmark sample is coethnic to their national leader.[17]

In Table 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 10.9 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 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. It is also evident from Panel B that there is valuable information in relative similarities that is not observable in a binary framework. While I reserve more conclusive statements for the regression analysis, this suggests that there is value added from using a continuous measure of similarity to study ethnic favoritism. Together these results clearly 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 cor-

[17]Table A6 reports a complete set of descriptive statistics used throughout this analysis.

Figure 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.

relation of 0.32 between light intensity and lexicostatistical similarity is also suggestive of this positive relationship (correlations not shown here).

I also plot average luminosity before and after a leadership change in Figure 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 large jump 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). All together, Figure 5 is a clean visualization of favoritism across linguistic lines.[18]

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 their 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:

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

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 $y_{c,l,t} \equiv \ln(0.01 + NightLights_{c,l,t})$.

$LS_{c,l,t-1}$, 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,[19] and the (logged) geodesic distance be-

[18] 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.

[19] This 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.

tween language group l and the language group associated with the leader of country c. 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.[20] $\gamma_{c,l}$ are country-language group fixed effects, $\lambda_{c,t}$ are country-year fixed effects and $\theta_{l,t}$ are language-year fixed effects.[21] In all specifications I adjust standard errors for clustering in country-language groups.[22]

3.1 Identification of Linguistic Similarity

In order to identify the effect of linguistic similarity it's 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. It's the arbitrary nature of border formation that forms the basis of my identification strategy.

The use of straight lines prevailed when drawing borders in African 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 nearly 80% of all

[20] See Appendix A for more details on data definitions and sources.

[21] In my benchmark sample $\gamma_{c,l}$ represents 357 fixed effects, $\lambda_{c,t}$ represents 680 fixed effects and $\theta_{l,t}$ represents 2959 fixed effects. Given the high-dimensionality of this triple difference fixed effects estimator, I use the Stata command `reghdfe` to estimate equation (1), a generalization of the standard `areg`/`xtreg` commands. Written by Sergio Correia, this estimator is based on the fixed point iteration of Guimarães and Portugal (2010) and Gaure (2013).

[22] 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.

African borders follow lines of latitude and longitude, an amount larger than any other continent in the world (Alesina et al., 2011) (see Englebert et al. (2002) and Michalopoulos and Papaioannou (2014, 2016) for a detailed discussion of the arbitrariness of African borders).

In this paper I study language groups partitioned across borders because the ethnolinguistic identity of a national leader varies by country. Hence the partitioning of a group generates within-group variation in terms of that group's linguistic similarity to their leader. The exogeneity of this variation comes from the arbitrary placement of borders established during the colonial annexation of Africa.

This strategy is similar to Michalopoulos and Papaioannou (2014), though a key difference is that I possess time-variation so the relative similarity within a partitioned group also varies over time as new leaders come to power from different ethnolinguistic groups. 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. Because I observe a partitioned group in at least two countries in each year, 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) I am estimating the effect of linguistic similarity off of changes in the incoming leader's ethnolinguistic identity.

4 Benchmark Results

I report estimates for 9 different versions of equation (1) for each of the 3 linguistic similarity measures in Table 3. The structure of the table is as follows: columns (1)-(3) report estimates with country-year fixed effects, columns (4)-(6) add country-language fixed effects to the estimates and columns (7)-(9) report estimates for the triple difference-in-differences estimator – my preferred specification. 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 3. Hereafter I will refer to column (9) as my benchmark specification.

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. To give economic meaning to these estimates, consider the benchmark estimate of lexicostatistical similarity in column (9). Using the rule of thumb that the estimated elasticity of GDP with respect to night lights is 0.3 (Henderson et al., 2012), the point estimate of 0.413 implies that a standard deviation increase in linguistic similarity (22.9 percent change) yields a 2.8 percent increase in regional GDP, an economically significant effect.[23]

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.

In Table 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

[23] $\%\Delta GDP_{c,l,t} \approx \%\Delta NightLights_{c,l,t} \times 0.3 = (\beta \times \Delta LS_{c,l,t-j}) \times 0.3 = 0.413 \times 0.229 \times 0.3 = 2.8\%$, assuming that $ln(0.01 + NightLights_{c,l,t}) \approx ln(NightLights_{c,l,t})$.

Table 3: Benchmark Regressions Using Various Measures of Linguistic Similarity

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	1.413***	0.914***	1.128***	0.164*	0.247**	0.272***	0.248*	0.376***	0.413***
	(0.274)	(0.328)	(0.338)	(0.093)	(0.102)	(0.101)	(0.139)	(0.143)	(0.141)
Adjusted R^2	0.340	0.426	0.452	0.919	0.919	0.919	0.923	0.924	0.924
Cladistic similarity$_{t-1}$	0.891***	0.458**	0.426**	0.088	0.096	0.084	0.267*	0.292**	0.266*
	(0.215)	(0.215)	(0.210)	(0.086)	(0.084)	(0.083)	(0.138)	(0.138)	(0.140)
Adjusted R^2	0.325	0.424	0.446	0.919	0.919	0.919	0.923	0.924	0.924
Coethnic$_{t-1}$	1.191***	0.558	0.907***	0.148*	0.261***	0.271***	0.094	0.162	0.220*
	(0.261)	(0.352)	(0.340)	(0.076)	(0.091)	(0.089)	(0.132)	(0.128)	(0.126)
Adjusted R^2	0.330	0.421	0.446	0.919	0.919	0.919	0.923	0.923	0.924
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	357	357	357	357	357	357	357	357	357
Countries	35	35	35	35	35	35	35	35	35
Language groups	164	164	164	164	164	164	164	164	164
Observations	6,475	6,475	6,475	6,475	6,475	6,475	6,475	6,475	6,475

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.

the fact that the effect of coethnicity is not identifiable in these regressions. In column (3) cladistic similarity outperforms coethnicity, 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$_{t-1}$ = (1 − Coethnic$_{t-1}$) × Lexicostatistical similarity$_{t-1}$, 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.

Consider non-coethnic lexicostatistical similarity in column (5). The observable variation of non-coethnic similarity is equivalent to the variation I identify off of in column (2), only now I've loaded 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, I find that coethnics are most favored with an estimated increase of 0.361 in average night light luminosity. While there is still an observable benefit from similarity among non-coethnics, the magnitude of the effect is 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 0.087 ($= 0.146 \times 0.596$) in night light luminosity.[24]

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

[24]By these estimates the threshold value of non-coethnic similarity is 0.605, 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 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$_{t-1}$	0.413**	0.596**		0.656**		
	(0.170)	(0.266)		(0.299)		
Cladistic similarity$_{t-1}$	0.000		0.219	-0.058		
	(0.163)		(0.159)	(0.161)		
Coethnic$_{t-1}$		-0.236	0.119	-0.255	0.361***	0.339**
		(0.241)	(0.143)	(0.243)	(0.130)	(0.150)
Non-coethnic lexicostatistical similarity$_{t-1}$					0.596**	
					(0.266)	
Non-coethnic cladistic similarity$_{t-1}$						0.219
						(0.159)
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	357	357	357	357	357	357
Countries	35	35	35	35	35	35
Language groups	164	164	164	164	164	164
Adjusted R^2	0.924	0.924	0.924	0.924	0.924	0.924
Observations	6,475	6,475	6,475	6,475	6,475	6,475

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 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.

Taken together the results of Table 3 and Table 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 5 reproduces the benchmark estimate of lexicostatistical similarity for comparison. In column (2) 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 (3) I report estimates from the joint estimation of lead and lagged lexicostatistical similarity. Again I find no evidence of a pre-trend effect in the lead variable, while the estimated effect of lagged lexicostatistical similarity is positive and significant.[25]

Next I re-estimate equation (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 (4). Lexicostatistical similarity is estimated to be positive and significant at a standard level of confidence, albeit with a reduced magnitude. Columns (2)-(4) are also consistent with Figure 5, which shows a lack of anticipatory changes in night lights preceding a change in leadership. Hence, these results are reassuring that

[25] I also examine the effect of linguistic similarity at deeper lags. I plot these coefficient estimates in Figure C1.

Table 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)
Lexicostatistical similarity$_{t-1}$	0.413***		0.388***	0.225***
	(0.141)		(0.130)	(0.082)
Lexicostatistical similarity$_{t+1}$		0.222	0.117	
		(0.136)	(0.114)	
Night lights$_{t-1}$				0.502***
				(0.049)
Geographic controls	Yes	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes
Clusters	357	357	357	357
Countries	35	35	35	35
Language groups	164	164	164	164
Adjusted R^2	0.924	0.933	0.934	0.945
Observations	6,475	5,955	5,955	6,178

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 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.

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 neighbouring 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 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 6. If people are in fact migrating in response to changes in their groups similarity, I should observe corresponding changes in population density. In all specifications, the various measures of similarity are insignificant, with the exception of the least squares estimate for coethnicity in column (7). However, this estimate is only indicative of correlation, and does not account for factors such as group size, the legacy of a group in a country, etc. When accounting for these country-group features the estimate goes to zero, as column (8) and (9) indicate. Overall these estimates imply that changes in night lights within a partitioned group cannot be explained away by movements of people to regions that are similar to the leader in terms of ethnolinguistic identity.

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 C. In particular, I show that the results are similar when:

Table 6: Test for Cross-Border Migration Following Leadership Changes

Dependent Variable: ln(Population Density$_{c,l,t}$)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.075	0.006	-0.014						
	(0.295)	(0.054)	(0.019)						
Cladistic similarity$_{t-1}$				-0.011	-0.053	-0.053			
				(0.239)	(0.044)	(0.044)			
Coethnic$_{t-1}$							0.613**	0.014	-0.013
							(0.302)	(0.043)	(0.016)
Country-language fixed effects	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Country-year fixed effects	No	No	Yes	No	No	Yes	No	No	Yes
Language-year fixed effects	No	No	Yes	No	No	Yes	No	No	Yes
Observations	6,475	6,475	6,475	6,475	6,475	6,475	6,475	6,475	6,475
Adjusted R^2	0.000	0.985	0.999	0.000	0.999	0.999	0.007	0.985	0.999
Clusters	357	357	357	357	357	357	357	357	357
Countries	35	35	35	35	35	35	35	35	35
Language groups	164	164	164	164	164	164	164	164	164

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 reproduce my benchmark estimates on a balanced panel of 79 ethnolinguistic groups partitioned across 22 countries (Table C1).
- I re-estimate equation (1) and weight the estimates by the Ethnologue population of each language group (Table C2). 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 I 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 C3).

4.1 The Dynamics of Incumbency

One drawback of estimating equation (1) is that I cannot account for long-run patterns of favoritism among leaders who hold office for many years. In this section I study the dynamics of my findings by accounting for the possibility that the extent of favoritism is a function of the time a leader has held office.

In column (1) of Table 7 I report estimates of an augmented equation (1) that includes an interaction between linguistic similarity and a leader's current years of incumbency. The interaction term enters positive and statistically significant, indicating that favoritism is an increasing function of the years a leader has held office. This introduced heterogeneity also dominates the direct effect of linguistic similarity, which itself loses significance after accounting for this interaction term.

In column (2) I report estimates of a similar model, but instead interact the total number of years a leader will hold office. Unlike the current years of incumbency, total years in office is constant across leaders. Nonetheless, I find a similar pattern: the extent of favoritism towards similar groups is increasing in the length of leadership, and when accounting for this heterogeneity the direct effect of linguistic similarity goes to zero.

Next I separate the current number of years each leader has held office by quartiles, and construct a series of dummy variables indicating the respective

Table 7: The Dynamics of Ethnolinguistic Favoritism

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$			
	(1)	(2)	(3)
Lexicostatistical similarity$_{t-1}$	0.132	0.111	
	(0.162)	(0.197)	
Lexicostatistical similarity$_{t-1}$	0.031**		
$\times$ Current years in office$_{t-1}$	(0.015)		
Lexicostatistical similarity$_{t-1}$		0.024*	
$\times$ Total years in office$_{t-1}$		(0.014)	
Lexicostatistical similarity$_{t-1}$			0.142
$\times$ 1(1^{st} quartile of years in office$_{t-1}$)			(0.157)
Lexicostatistical similarity$_{t-1}$			0.142
$\times$ 1(2^{nd} quartile of years in office$_{t-1}$)			(0.135)
Lexicostatistical similarity$_{t-1}$			0.517***
$\times$ 1(3^{rd} quartile of years in office$_{t-1}$)			(0.171)
Lexicostatistical similarity$_{t-1}$			0.611**
$\times$ 1(4^{th} quartile of years in office$_{t-1}$)			(0.241)
Geographic controls	Yes	Yes	Yes
Distance & population density	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes
Clusters	357	357	357
Countries	35	35	35
Language groups	164	164	164
Adjusted R^2	0.925	0.924	0.924
Observations	6,475	6,475	6,475

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 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.

quartile. I interact each of these four dummy variables with lexicostatistical similarity, and report the estimates in column (3). All coefficients are positive and the magnitude of effect is monotonically increasing in the length of tenure. Only the third and fourth quartile enter positive and significant, suggesting that the benchmark result is largely driven by observations where the incumbent leader holds power for longer than the sample median of 9 years.

Taken together Table 7 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. The policy implication is that term limits can be an effective tool to minimize the extent of favoritism throughout Africa.

5 Mechanism

In this section I explore two channels that can possibly explain the benchmark finding of ethnolinguistic favoritism. With conventional forms of favoritism such as road building, the linkage is straightforward: a leader invests in infrastructure in his/her preferred region because of the leader's ethnolinguistic connection to the region. I call this the locational mechanism because favoritism is based on the similarity of a leader to a particular region. Locational favoritism is beneficial to all inhabitants of the region regardless of their background. I also explore an individual-level mechanism. I ask whether the effects of linguistic similarity are beneficial irrespective of where an individual lives – i.e., do individuals who live outside of their ethnolinguistic homeland benefit from a similar linguistic background to their leader?

5.1 DHS Individual-Level Data

To separate these two channels I use data from the Demographic and Health Surveys (DHS) for 13 African countries.[26] For each country I pool both the

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

male and female samples for each wave, and when separately provided, I merge the wealth index dataset for that year. In an effort to replicate the same variation I use in my benchmark estimates, I choose DHS countries and survey waves in the following way:

(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.[27] 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 A. 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 groups allows for a set-up similar to my benchmark model, where variation in locational and individual similarity comes from leadership

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

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.

Of the 56,455 DHS survey respondents that I successfully match both measures of similarity, I find that 55.9 percent reside in their ethnolinguistic homeland.[28] 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. However, the respondents most important to this section are those that do not reside in their ethnolinguistic homeland (i.e., non-natives). By including both measures of similarity in a single estimating equation, I exploit variation in individual and locational similarity within non-native individuals to separately estimate the two effects off of changes in a leader's ethnolinguistic identity across waves.[29]

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.).

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), I include individual language-wave fixed effects because 45 percent of respondents' home language is different than their locational language.

[28] 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).

[29] 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.

Table 8 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.[30] The magnitude of locational similarity increases slightly and maintains its strong significant effect on individual wealth. In Table C4 I add each individual control variable one at a time. While I account for capital city effects with an indicator variable, I also account for additional spatial effects in columns (4) and (5) by separately adding distance to the nearest border and distance to the coast in levels.[31]

The middle panel of Table 8 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

[30]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 A for variable definitions.

[31]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.

Table 8: 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 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.

are consistent with the rest of the table and reported in the bottom panel of Table 8. 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.

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, 2014). 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 (2014) 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: non-coethnic locational similarity $=$ (1 $-$ coethnicity) $\times$ locational similarity. Table 9 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.

Table 9: 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***	0.601***	0.324**
	(0.430)	(0.139)	(0.116)	(0.160)	(0.134)
Non-coethnic locational similarity$_{t-1}$	-0.692	0.348*	0.697***	0.854***	0.573***
	(0.556)	(0.205)	(0.148)	(0.173)	(0.167)
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
Adjusted R^2	0.314	0.574	0.603	0.603	0.604
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6 Concluding Remarks

An important aspect of Africa's colonial history was the imposition of relatively few nation-states on top of a multitude of autonomous ethnic groups. The salience of ethnicity in African politics today can be traced back to this forced coexistence, where ethnicity became a vanguard of identity. Yet the camaraderie found among coethnics is also a means to discrimination among non-coethnics. The empirical literature has done a good job documenting the importance of coethnic relationships between leaders and citizens, where coethnics receive preferential treatment from their national leader because of their shared background. However the binary nature of a coethnic relationship neglects the fact that some non-coethnics are more similar to their leader than others. This neglect is, in part, due to the absence of a suitable measure of relative similarity.

In this paper, I introduce a lexicostatistical measure of linguistic similarity that, to its advantage, is a continuous measure of group relatedness. The novelty of this approach is that I can measure relative similarities among non-coethnics, something that is not possible in the binary world of coethnicity.

Using this new measure of similarity I find robust evidence of favoritism in 164 languages groups split across 35 African countries, a phenomenon I term ethnolinguistic favoritism. In the absence of data on political patronage at the language group level, I use night light luminosity to capture patterns of favoritism. I also document that there is valuable information in my continuous measure of similarity that is not observable with a measure of coethnicity. This evidence supports my hypothesis of favoritism working across a gradient of group similarity in Africa.

I also ask whether ethnolinguistic favoritism is purely a regional phenomenon, or whether individuals are better off the more linguistically similar they are to their leader irrespective of where they live. Contrary to anecdotal evidence on the benefits of individual similarity, the evidence I present here suggests that the benefits of favoritism are regionally distributed.

References

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

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.

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.

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

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

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

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.

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.

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

Dreher, A., Fuchs, A., Hodler, R., Parks, B. C., Raschky, P. A., and Tierney, M. J. (2014). Aid on Demand: African Leaders and the Geography of China's Foreign Assistance. *AidData Wokring Paper 3*, (November).

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.

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.

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.

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 Ozak, O. (2015). The Agricultural Origins of Time Preference. *Brown University, mimeo*, pages 1–107.

Gaure, S. (2013). OLS with Multiple High Dimensional Category Variables. *Computational Statistics and Data Analysis*, 66:8–18.

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

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

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.

Golden, M. and Min, B. (2013). Distributive Politics Around the World. *Annual Review of Political Science*, 16(1):73–99.

Gomes, J. F. (2014). The Health Costs of Ethnic Distance: Evidence from Sub-Saharan Africa. *University of Navarra, mimeo*.

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

Guimarães, P. and Portugal, P. (2010). A Simple Feasible Procedure to Fit Models with High-Dimensional Fixed Effects. *The Stata Journal*, 10(4):628–649.

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

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.

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.

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.

Kramon, E. and Posner, D. N. (2013). Who Benefits from Distributive Politics? How the Outcome One Studies Affects the Answer One Gets. *Perspectives on Politics*, 11(2):461–474.

Kramon, E. and Posner, D. N. (2014). Ethnic Favoritism in Primary Education in Kenya. *UCLA, mimeo*.

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

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*.

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

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

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.

Solon, G., Haider, S. J., and Wooldridge, J. (2013). 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. (2013). How Deep Are the Roots of Economic Development? *Journal of Economic Literature*, 51(2):325–369.

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*, (Forthcoming).

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.

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.

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.

FOR ONLINE PUBLICATION

A Data Descriptions, Sources and Summary Statistics

A.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: Linguistic similarity is constructed using two different measures of 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 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. (2014), and 2012-2013 using Wikipedia.
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 Ozak (2015). 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/

A.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/

Individuall 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 check 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 then 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 then measure the geodesic distance between the two points in Stata using the `globdist` command.
Source: http://www.worldgeodatasets.com/language/

A.3 Summary Statistics and Additional Details

Table A1: 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 A2: 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 A3: 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 A4: 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 A5: 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 A6: Summary Statistics – Regional-Level Dataset

	Mean	Std dev.	Min	Max	N
Night lights	0.122	0.386	0.000	4.540	6,475
$\ln(0.01 + \text{night lights}_t)$	-3.496	1.423	-4.605	1.515	6,475
$\ln(0.01 + \text{night lights}_{t-1})$	-3.514	1.412	-4.605	1.515	6,178
Lexicostatistical similarity$_{t-1}$	0.192	0.229	0.000	1.000	6,475
Cladistic similarity$_{t-1}$	0.411	0.327	0.000	1.000	6,475
Coethnicity$_{t-1}$	0.046	0.210	0.000	1.000	6,475
Non-coethnic cladistic similarity$_{t-1}$	0.365	0.310	0.000	0.966	6,475
Non-coethnic lexicostatistical similarity$_{t-1}$	0.146	0.148	0.000	0.960	6,475
Lexicostatistical similarity$_{t+1}$	0.193	0.229	0.000	1.000	6,070
Current years in office$_{t-1}$	11.54	8.786	1.000	38.00	6,475
Total years in office$_{t-1}$	18.63	10.29	1.000	38.00	6,475
Log distance (km) to leader's group$_{t-1}$	5.839	1.475	0.000	7.419	6,475
Log population density$_t$	2.920	1.487	-2.169	6.116	6,475
Absolute difference in elevation$_t$	263.8	330.8	0.000	2,581	6,475
Absolute difference in ruggedness$_t$	104.1	109.9	0.000	548.8	6,475
Absolute difference in precipitation$_t$	31.17	29.61	0.000	230.7	6,475
Absolute difference in mean temperature$_t$	17.42	18.34	0.000	130.2	6,475
Absolute difference in caloric suitability index$_t$	299.7	311.0	0.000	1,711	6,475
Oil reserve in both leader and language group$_t$	0.018	0.134	0.000	1.000	6,475
Diamond mine in both leader and language group$_t$	0.079	0.270	0.000	1.000	6,475

Table A7: 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

B Measures of Linguistic Similarity

B.1 Computerized Lexicostatistical Similarity

The computerized approach to estimating lexicostatistical distances was developed as part of the *Automatic Similarity Judgement Program* (ASJP), a project run by linguists at the Max Planck Institute for Evolutionary Anthropology. 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.[32]

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

I use a variant of the Levenshtein distance algorithm, which in its simplest form calculates the minimum number of edits necessary to translate the spelling of a word from one language to another. In particular, I use the normalized and divided Levenshtein distance estimator proposed by Bakker et al. (2009).[33] 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 $1 \leq i \leq M$.[34] 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

[32] 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).

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

[34] 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:

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

where $L(\alpha_i, \beta_i)$ is the length of the longer of the two spellings α_i and β_i of word i. $LDN(\alpha_i, \beta_i)$ is the normalized Levenshtein distance, which represents a percentage estimate of dissimilarity between languages α and β for word i. For each language pair, $LDN(\alpha_i, \beta_i)$ is calculated for each word of the M-word Swadesh list. Then the average lexical distance for each language pair is calculated by averaging across all M words for those two languages. The average distance between two languages is then

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

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

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

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

implemented by weighting equation (3) with equation (4) as follows:

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

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

$$LS(\alpha, \beta) = 1 - LDND(\alpha, \beta). \tag{6}$$

B.2 Cladistic Similarity

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

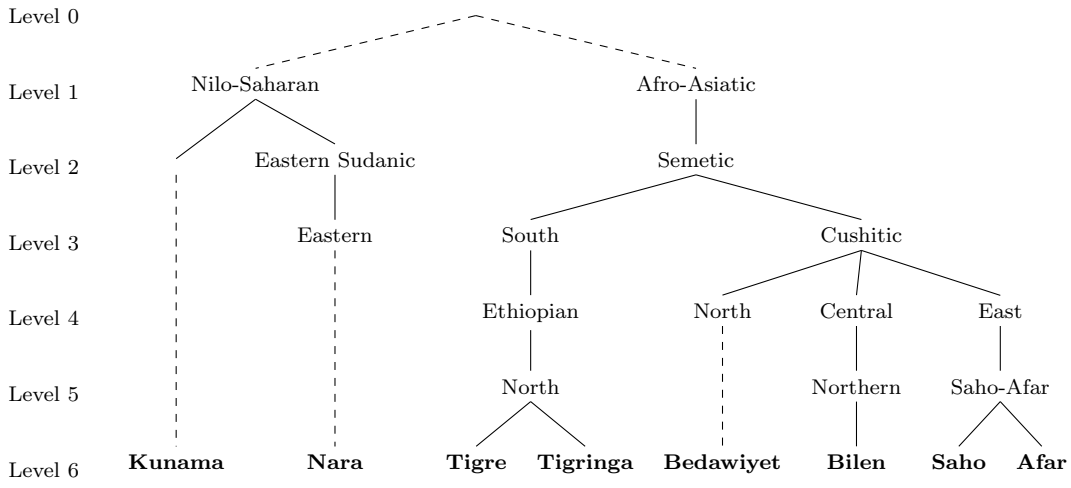
$$CS(\alpha, \beta) = \left(\frac{s(\alpha, \beta)}{M} \right)^{\delta}, \tag{7}$$

where δ is an arbitrarily assigned weight used to discount more recent linguistic cleavages relative to deep cleavages. I describe this weight as arbitrary because there is no consensus on the appropriate weight to be assumed. Fearon (2003) argues the true function is probably 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 (not shown here).

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

Figure B1: Phylogenetic Tree of Eritrean Languages



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

C Supplementary Material

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

C.1 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 sample to the 79 partitioned language groups for which I observe lexicostatistical similarity in every year. Table C1 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 3.

C.2 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., 2013). Table C2 reports these estimates.

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.

C.3 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 C3.

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.

C.4 DHS Controls

In this section I 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 C4 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.

Table C1: Robustness Check: Benchmark Regressions on a Balanced Panel

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	1.310***	1.057**	1.504***	0.370**	0.408**	0.442***	0.324	0.468*	0.503**
	(0.473)	(0.492)	(0.514)	(0.156)	(0.160)	(0.160)	(0.260)	(0.247)	(0.253)
Adjusted R^2	0.384	0.473	0.516	0.933	0.934	0.934	0.937	0.937	0.938
Cladistic similarity$_{t-1}$	0.848**	0.468	0.421	0.265	0.201	0.208	0.398	0.436	0.386
	(0.420)	(0.417)	(0.407)	(0.240)	(0.229)	(0.226)	(0.324)	(0.320)	(0.330)
Adjusted R^2	0.373	0.469	0.505	0.933	0.934	0.934	0.937	0.937	0.938
Coethnic$_{t-1}$	0.928*	0.332	0.770	0.253**	0.327**	0.338**	0.047	0.078	0.130
	(0.481)	(0.567)	(0.511)	(0.121)	(0.148)	(0.148)	(0.266)	(0.248)	(0.251)
Adjusted R^2	0.370	0.466	0.505	0.933	0.934	0.934	0.936	0.937	0.937
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	167	167	167	167	167	167	167	167	167
Countries	22	22	22	22	22	22	22	22	22
Language groups	79	79	79	79	79	79	79	79	79
Observations	3,674	3,674	3,674	3,674	3,674	3,674	3,674	3,674	3,674

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 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 C2: Robustness Check: Benchmark Regressions Weighted by Language Group Population

Dependent Variable: $y_{c,l,t} = \ln(0.01 + NightLights_{c,l,t})$									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	1.164***	0.657***	0.852***	0.032	0.060	0.076	0.264	0.445**	0.450**
	(0.328)	(0.236)	(0.251)	(0.046)	(0.057)	(0.061)	(0.162)	(0.206)	(0.199)
Adjusted R^2	0.730	0.827	0.844	0.983	0.983	0.984	0.990	0.990	0.990
Cladistic similarity$_{t-1}$	1.164***	0.353	0.314	-0.016	-0.020	-0.007	0.322	0.390*	0.368*
	(0.315)	(0.298)	(0.267)	(0.047)	(0.045)	(0.044)	(0.199)	(0.212)	(0.208)
Adjusted R^2	0.723	0.825	0.840	0.983	0.983	0.984	0.990	0.990	0.990
Coethnic$_{t-1}$	0.710**	0.452*	0.462*	0.007	0.051	0.054	0.165	0.321**	0.362***
	(0.348)	(0.234)	(0.246)	(0.026)	(0.033)	(0.033)	(0.127)	(0.129)	(0.137)
Adjusted R^2	0.715	0.825	0.840	0.983	0.983	0.984	0.990	0.990	0.990
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	357	357	357	357	357	357	357	357	357
Countries	35	35	35	35	35	35	35	35	35
Language groups	164	164	164	164	164	164	164	164	164
Observations	6,475	6,475	6,475	6,475	6,475	6,475	6,475	6,475	6,475

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 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 C3: Robustness Check: Alternative Dependent Variables

	(1)	(2)	(3)	(4)	(5)	(6)
Lexicostatistical similarity$_{t-1}$	0.055***			0.476*		
	(0.020)			(0.283)		
Cladistic similarity$_{t-1}$		0.041**			0.334	
		(0.020)			(0.317)	
Coethnic$_{t-1}$			0.026			0.194
			(0.016)			(0.190)
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	357	357	357	357	357	357
Countries	35	35	35	33	33	33
Language groups	164	164	164	100	100	100
Adjusted R^2	0.953	0.953	0.953	0.929	0.929	0.929
Observations	6,475	6,475	6,475	2,769	2,769	2,769

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 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 C4: Individual-Level Regressions: Baseline Covariates

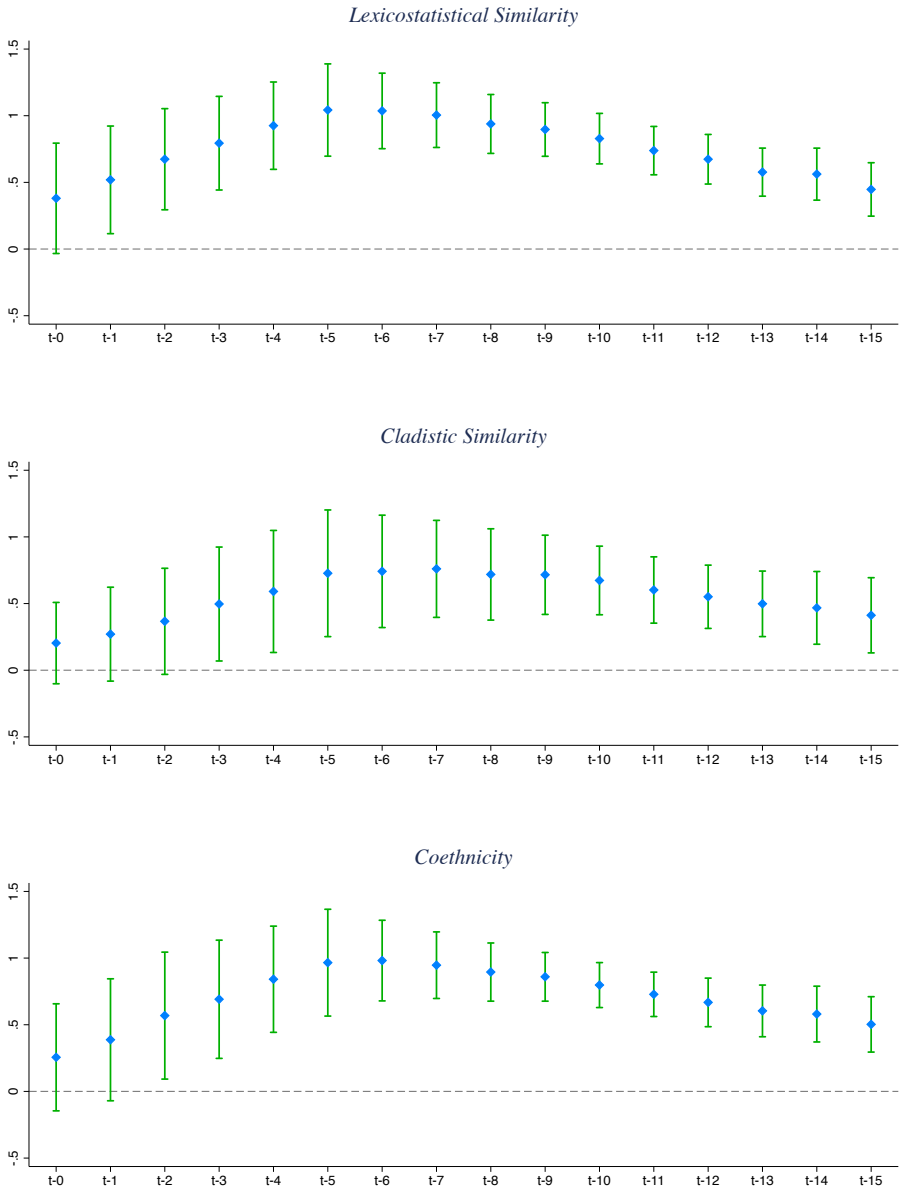
Dependent Variable: DHS Wealth Index									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Locational similarity$_{t-1}$	0.585	0.594	0.463***	0.636	0.490	1.024*	0.518	0.608	0.479***
	(0.604)	(0.613)	(0.152)	(0.398)	(0.637)	(0.592)	(0.587)	(0.399)	(0.119)
Age	-0.021***								-0.008
	(0.005)								(0.006)
Age 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 5. The unit of observation is an individual. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C.3 The Lagged Effect of Linguistic Similarity

Throughout the empirical analysis of this paper I use a single period lag when measuring linguistic similarity to account for an expected delay between a leader coming to power and the legislation of policy that favors his/her preferred group(s). In this section I extend my data on African leaders back to 1977 using the same methodology outlined in section 2.3. I then re-estimate equation (1), adjusting the lagged value of linguistic similarity one period at a time from $j = 0, \ldots, 15$. I plot these point estimates for all three measure of linguistic similarity, including their 95 percent confidence interval, in Figure C1. For lexicostatistical similarity, the estimate increases monotonically up to $j - 5$ and begins to revert towards zero thereafter. This hump-shaped time pattern is consistent across each alternative measure, although lexicostatistical similarity continues to out perform cladistic similarity and coethnicity in terms of precision. Furthermore, the point estimate is increasing in magnitude and significance between $j-1$ – the one-period lag used in the benchmark model – and $j - 5$ which suggests that, if anything, the benchmark findings are conservatively estimated.

Figure C1: The Effect of Linguistic Similarity on Night Light Intensity



These figures display the effect of three measures of linguistic similarity on night light intensity. In each figure I plot a series of estimates where the lagged value of linguistic similarity is lagged one period at a time from $j = 0, \ldots, 15$. Night light intensity is measured in language group l of country c for year t, and linguistic similarity measures the similarity between group l and the leader of country c in year $t - j$, where $j = 0, \ldots, 15$. All regressions include country-year fixed effects, language-year effects, country-language fixed effects, and log population density. The observations vary by the depth of the lag, from 4,862 observations in $t - 0$ to 4,512 observations in $t - 15$. Intervals reflect 95% confidence levels.