Ethnolinguistic Favoritism in African Politics
ONLINE APPENDIX

Andrew Dickens[†]

For publication in the
American Economic Journal: Applied Economics

[†]Brock University, Department of Economics, 1812 Sir Issac Brock Way, L2S 3A2, St. Catharines, ON, Canada (email: adickens@brocku.ca).

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: I construct two measures of linguistic similarity: lexicostatistical similarity from the Automatic Similarity Judgement Program (ASJP), and cladistic similarity using Ethnologue data from the WLMS. I use these to measure the similarity between each language group and the ethnolinguistic identity of that country's national leader. I discuss how I assign a leader's ethnolinguistic identity in Section 1 of the paper.
Source: http://asjp.clld.org and http://www.worldgeodatasets.com/language/

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

Population density: Population density is calculated by averaging across pixels that fall within each country-language group polygon. To minimize area distortions I use the Africa Albers Equal Area Conic projection. Data comes from the Gridded Population of the World, which is available in 5-year intervals: 1990, 1995, 2000, 2005, 2010. For intermediate years I assume population density is constant; e.g., the 1995 population density is assigned to years 1995-1999. Throughout the regression analysis I use log population density.
Source: http://sedac.ciesin.columbia.edu/data/collection/gpw-v3

National leaders: I collected birthplace locations of all African leaders between 1991-2013. Names of African leaders and years entered and exited office comes from the Archigos

Database on Leaders 1875-2004 (Goemans et al., 2009), which I extended to 2011 using data from Dreher et al. (2015), and 2012-2013 using a country's Historical Dictionary and other secondary sources.
Source: http://www.rochester.edu/college/faculty/hgoemans/data.htm

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

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

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

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

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

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

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

Absolute difference in caloric suitability index: I sourced the caloric suitability index (CSI) data from Galor and Ozak (2016). CSI is a measure of agricultural productivity that reflects the caloric potential in a grid cell. It's based on the Global Agro-Ecological Zones (GAEZ) project of the Food and Agriculture Organization (FAO). A variety of related measures are available: in the reported estimates I use the pre-1500 average CSI measure that includes cells with zero productivity. The results are not sensitive to which measure I use. I measure average CSI within each partitioned language group and leader's ethnolinguistic group using the Africa Albers Equal Area Conic projection. I use Stata to calculate the absolute difference between the two.
Source: http://omerozak.com/csi

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

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

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/

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

1. Direct match: the DHS ethnicity name is the same as an Ethnologue language name for the respondent's country of residence.

2. Alternative name: the unmatched DHS ethnicity is an unambiguous alternative name for a language in the Ethnologue or Glottolog database.

3. Macrolanguage: if the ethnicity corresponds to a macrolanguage in the Ethnologue, then I assign the most populated sub-language of that macrolanguage.

4. Population size: if the unmatched ethnicity maps to numerous languages, I choose the language with the largest Ethnologue population.

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

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

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

Wealth Index for more details.

Age: Age of respondent at the time of survey.

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

Rural: An indicator variable for rural locations.

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

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

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

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

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

A.3 Summary Statistics and Additional Details

Table A1: Summary Statistics – Regional-Level Dataset

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

Table A2: Summary Statistics – DHS Individual-Level Dataset

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

Table A3: Summary Statistics – Power Sharing Dataset

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

Table A4: Leadership by Country – Regional-Level Dataset

Country	Leader Name	Entered Office	Left Office	Ethnolinguistic Group	Sample Years
Angola	Jose Eduardo dos Santos	1979	Ongoing	Kimbundu	1992-2013
Benin	Mathieu Kerekou	1996	2006	Waama	1996-2006
Botswana	Quett Masire	1980	1998	Tswana	1992-1998
Botswana	Festus Mogae	1998	2008	Kalanga	1999-2008
Burkina Faso	Blaise Compaore	1987	Ongoing	Moore	1992-2013
Cameroon	Paul Biya	1982	2013	Bulu	1992-2013
Central African Republic	Andre-Dieudonne Kolingba	1981	1993	Yakoma	1992-1993
Central African Republic	Ange-Felix Patasse	1993	2003	Kaba	1994-2003
Chad	Idriss Deby	1990	Ongoing	Zaghawa	1992-2013
Congo	Denis Sassou Nguesso	1979	1992	Mbosi	1992
Congo	Pascal Lissouba	1992	1997	Punu	1993-1997
Congo	Denis Sassou Nguesso	1997	Ongoing	Mbosi	1998-2013
Cote d'Ivoire	Houphouet-Boigny	1960	1993	Baoule	1992-1993
Cote d'Ivoire	Konan Bedie	1993	1999	Baoule	1994-1999
Cote d'Ivoire	Robert Guei	1999	2000	Dan	2000
Cote d'Ivoire	Alassane Ouattara	2011	Ongoing	Jula	2012-2013
DRC	Mobutu Sese Seko	1965	1997	Ngbandi, Southern	1992-1997
DRC	Laurent-Desire Kabila	1997	2001	Luba-Kasai	1998-2001
DRC	Joseph Kabila	2001	Ongoing	Luba-Kasai	2002-2013
Eritrea	Isaias Afewerki	1993	Ongoing	Tigrigna	1994-2013
Ethiopia	Meles Zenawi	1991	2012	Tigrigna	1992-2012
Ethiopia	Hailemariam Desalegn	2012	2013	Wolaytta	2013
Gambia	Dawda Jawara	1965	1994	Mandinka	1992-1994
Gambia	Yahya Jammeh	1994	Ongoing	Jola-Fonyi	1995-2013
Ghana	Jerry Rawlings	1981	2001	Ewe	1992-2001
Ghana	John Agyekum Kufuor	2001	2009	Akan	2002-2009
Ghana	John Evans Atta-Mills	2009	2012	Akan	2010-2012
Ghana	John Dramani Mahama	2012	Ongoing	Gonja	2013
Guinea	Lansana Conte	1984	2008	Susu	1992-2008
Guinea	Moussa Dadis Camara	2008	2009	Kpelle, Guinea	2009
Guinea-Bissau	Joao Bernardo Vieira	1980	1999	Papel	1992-1999
Guinea-Bissau	Malam Bacai Sanha	1999	2000	Mandinka	2000
Guinea-Bissau	Kumba Iala	2000	2003	Balanta-Kentohe	2001-2003
Guinea-Bissau	Henrique Pereira Rosa	2003	2005	Balanta-Kentohe	2004-2005
Guinea-Bissau	Joao Bernardo Vieira	2005	2009	Papel	2006-2009
Guinea-Bissau	Malam Bacai Sanha	2009	2012	Mandinka	2010-2012
Guinea-Bissau	Manuel Serifo Nhamadjo	2012	Ongoing	Pulaar	2013
Kenya	Daniel arap Moi	1978	2002	Tugen	1992-2002
Kenya	Mwai Kibaki	2002	2013	Gikuyu	2003-2013
Lesotho	Elias Phisoana Ramaema	1991	1993	Sotho, Southern	1992-1993
Lesotho	Ntsu Mokhehle	1993	1998	Sotho, Southern	1994-1998
Lesotho	Pakalithal Mosisili	1998	2012	Sotho, Southern	1999-2012
Lesotho	Tom Thabane	2012	Ongoing	Sotho, Southern	2013
Liberia	Wilton Sankawulo	1995	1996	Kpelle, Liberia	1995
Liberia	Ruth Perry	1996	1997	Vai	1996-1997
Liberia	Charles Taylor	1997	2003	Gola	1998-2003
Liberia	Ellen Johnson Sirleaf	2006	Ongoing	Gola	2007-2013
Malawi	Hastings Banda	1964	1994	Nyanja	1992-1994
Malawi	Bakili Muluzi	1994	2004	Yao	1995-2004

Malawi	Joyce Banda	2012	Ongoing	Tumbuka	2013
Mali	Amadou Toumani Toure	1991	1992	Bozo, Jenaama	1992
Mali	Alpha Oumar Konare	1992	2002	Pulaar	1993-2002
Mali	Amadou Toumani Toure	2002	2012	Bozo, Jenaama	2003-2012
Mali	Dioncounda Traore	2012	2013	Bamanankan	2013
Mozambique	Joaquim Alberto Chissano	1986	2005	Tsonga	1992-2005
Mozambique	Armando Emilio Guebuza	2005	Ongoing	Makhuwa	2006-2013
Namibia	Sam Daniel Nujoma	1990	2005	Ndonga	1992-2005
Namibia	Hifikepunye Pohamba	2005	Ongoing	Ndonga	2006-2013
Niger	Ali Saibou	1987	1993	Zarma	1992-1993
Niger	Mahamane Ousmane	1993	1996	Kanuri, Magna	1994-1996
Niger	Ibrahim Bare Mainassara	1996	1999	Hausa	1997-1999
Niger	Mamadou Tandja	1999	2010	Kanuri, Central	2000-2010
Niger	Mahamadou Issoufou	2011	Ongoing	Hausa	2011-2013
Nigeria	Sani Abacha	1993	1998	Kanuri, Central	1994-1998
Nigeria	Abdulsalami Abubakar	1998	1999	Gbagyi	1999
Nigeria	Olusegun Obasanjo	1999	2007	Yoruba	2000-2007
Nigeria	Umaru Musa Yar'Adua	2007	2010	Fulfulde, Nigerian	2008-2010
Senegal	Abdou Diouf	1981	2000	Serer-Sine	1992-2000
Senegal	Abdoulaye Wade	2000	2012	Wolof	2001-2012
Senegal	Macky Sall	2012	Ongoing	Serer-Sine	2013
Sierra Leone	Joseph Saidu Momoh	1985	1992	Limba, East	1992
Sierra Leone	Valentine Strasser	1992	1996	Krio	1993-1996
Sierra Leone	Ahmad Tejan Kabbah	1996	1997	Mende	1997
Sierra Leone	Johnny Paul Koroma	1997	1998	Limba, East	1998
Sierra Leone	Ahmad Tejan Kabbah	1998	2007	Mende	1999-2007
Sierra Leone	Ernest Bai Koroma	2007	Ongoing	Themne	2008-2013
Somalia	Ali Mahdi Muhammad	1991	1997	Somali	1992-1997
Somalia	Abdiqasim Salad Hassan	2000	2004	Somali	2000-2004
Somalia	Abdullahi Yusuf Ahmed	2004	2008	Somali	2005-2008
Somalia	Sharif Sheikh Ahmed	2009	2012	Somali	2009-2012
Somalia	Hassan Sheikh Mohamud	2012	Ongoing	Somali	2013
South Africa	F. W. de Klerk	1989	1994	Afrikaans	1992-1994
South Africa	Nelson Mandela	1994	1999	Xhosa	1995-1999
South Africa	Thabo Mbeki	1999	2008	Xhosa	2000-2009
South Africa	Jacob Zuma	2009	Ongoing	Zulu	2010-2013
Sudan	Omar al-Bashir	1989	Ongoing	Arabic, Sudanese	1992-2013
Tanzania	Ali Hassan Mwinyi	1985	1995	Zaramo	1992-1995
Tanzania	Jakaya Kikwete	2005	Ongoing	Kwere	2006-2013
Togo	Gnassingbe Eyadema	1967	2005	Kabiye	1992-2005
Togo	Faure Gnassingbe	2005	Ongoing	Kabiye	2006-2013
Uganda	Yoweri Museveni	1986	Ongoing	Nyankore	1992-2013
Zambia	Frederick Chiluba	1991	2002	Lamba	1992-2002
Zambia	Levy Mwanawasa	2002	2008	Lenje	2003-2008
Zambia	Michael Sata	2011	Ongoing	Bemba	2012-2013
Zimbabwe	Robert Mugabe	1980	Ongoing	Shona	1992-2013

Table A5: Leadership by Country – Individual-Level Dataset

Country	Leader Name	Entered Office	Left Office	Ethnolinguistic Group	DHS Survey Wave
Burkina Faso	Blaise Compaore	1987	Ongoing	Moore	2, 3, 4
DRC	Joseph Kabila	2001	Ongoing	Luba-Kasai	5, 6
Ethiopia	Meles Zenawi	1991	2012	Tigrigna	4, 5, 6
Ghana	Jerry Rawlings	1981	2001	Ewe	2, 3
Ghana	John Agyekum Kufuor	2001	2009	Akan	4
Guinea	Lansana Conte	1984	2008	Susu	4, 5
Guinea	Alpha Conde	2010	2013	Susu	6
Kenya	Mwai Kibaki	2002	2013	Gikuyu	4, 5, 6
Liberia	Ellen Johnson Sirleaf	2006	2013	Gola	5, 6
Mali	Alpha Oumar Konare	1992	2002	Pulaar	3, 4
Mali	Amadou Toumani Toure	2002	2012	Bozo, Jenaama	5, 6
Namibia	Hifikepunye Pohamba	2005	2013	Kwanyama	5, 6
Senegal	Abdou Diouf	1981	2000	Wolof	3
Senegal	Abdoulaye Wade	2000	2012	Wolof	4
Sierra Leone	Ernest Bai Koroma	2007	2013	Themne	5, 6
Uganda	Yoweri Museveni	1986	2013	Nyankore	5, 6
Zambia	Levy Mwanawasa	2002	2008	Lenje	5
Zambia	Michael Sata	2011	2013	Bemba	6

Table A6: Leadership by Country – Power Sharing Dataset

Country	Leader Name	Entered Office	Left Office	Ethnolinguistic Group	Sample Years
Benin	Mathieu Kerekou	1996	2006	Waama	1996-2004
Cameroon	Paul Biya	1982	2013	Bulu	1992-2004
Congo	Pascal Lissouba	1992	1997	Punu	1992-1996
Congo	Denis Sassou Nguesso	1997	2013	Mbosi	1997-2004
Cote d'Ivoire	Houphouet-Boigny	1960	1993	Baoule	1992
Cote d'Ivoire	Konan Bedie	1993	1999	Baoule	1993-1998
Cote d'Ivoire	Robert Guei	1999	2000	Dan	1999
Cote d'Ivoire	Laurent Gbagbo	2000	2011	Bete, Gagnoa	2000-2004
DRC	Mobutu Sese Seko	1965	1997	Ngbandi, Southern	1992-1996
DRC	Laurent-Desire Kabila	1997	2001	Luba-Kasai	1997-2000
DRC	Joseph Kabila	2001	2013	Luba-Kasai	2001-2004
Gabon	Omar Bongo Ondimba	1967	2009	Teke, Northern	1992-2004
Ghana	Jerry Rawlings	1981	2001	Ewe	1992-2000
Ghana	John Agyekum Kufuor	2001	2009	Akan	2001-2004
Guinea	Lansana Conte	1984	2008	Susu	1992-2004
Kenya	Daniel arap Moi	1978	2002	Tugen	1992-2001
Kenya	Mwai Kibaki	2002	2013	Gikuyu	2002-2004
Liberia	Amos Sawyer	1990	1994	Liberian English	1992-1993
Liberia	David Kpormapkor	1994	1995	Gola	1994
Liberia	Wilton Sankawulo	1995	1996	Kpelle, Liberia	1995
Liberia	Ruth Perry	1996	1997	Vai	1996
Liberia	Charles Taylor	1997	2003	Gola	1997-2002
Liberia	Gyude Bryant	2003	2006	Liberian English	2003-2004
Nigeria	Ibrahim Babangida	1985	1993	Gbagyi	1992
Nigeria	Sani Abacha	1993	1998	Kanuri, Central	1993-1997
Nigeria	Abdulsalami Abubakar	1998	1999	Gbagyi	1998
Nigeria	Olusegun Obasanjo	1999	2007	Yoruba	1999-2004
Sierra Leone	Valentine Strasser	1992	1996	Krio	1992-1995
Sierra Leone	Ahmad Tejan Kabbah	1996	1997	Mende	1996
Sierra Leone	Johnny Paul Koroma	1997	1998	Limba, East	1997
Sierra Leone	Ahmad Tejan Kabbah	1998	2007	Mende	1998-2004
Tanzania	Ali Hassan Mwinyi	1985	1995	Zaramo	1992-1994
Tanzania	Benjamin Mkapa	1995	2005	Makhuwa-Meetto	1995-2004
Togo	Gnassingbe Eyadema	1967	2005	Kabiye	1992-2004
Uganda	Yoweri Museveni	1986	2013	Nyankore	1992-2004

Table A7: 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 A8: 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 A9: 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

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.[1]

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).[2] Denote $LD(\alpha_i, \beta_i)$ as the raw Levenshtein distance for word i of languages α and β. Each word i comes from the aforementioned Swadesh list. Define the length of this list be M, so $1 \leq i \leq M$.[3] The algorithm is run to calculate $LD(\alpha_i, \beta_i)$ for each word in the M-word Swadesh list across each language pair. To correct for the fact that longer words will often demand more edits, the distance is normalized according to word length:

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

where $L(\alpha_i, \beta_i)$ is the length of the longer of the two spellings α_i and β_i of word i. $LDN(\alpha_i, \beta_i)$ is the normalized Levenshtein distance, which represents a percentage estimate of dissimilarity between languages α and β for word i. For each language pair, $LDN(\alpha_i, \beta_i)$ is calculated

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

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

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

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

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

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

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

where $GD(\alpha, \beta)$ is the global (average) distance between two languages excluding all word comparisons of the same meaning. This estimates the similarity of languages α and β only in terms of the ordering and frequency of characters, and is independent of meaning. The second normalization procedure is then implemented by weighting equation (2) with equation (3) as follows:

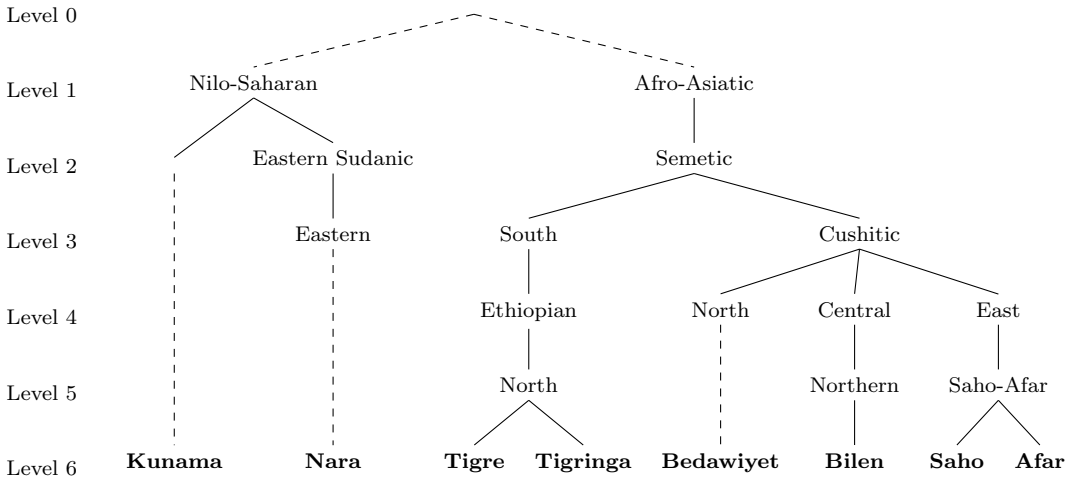
$$LDND(\alpha, \beta) = \frac{LDN(\alpha, \beta)}{GD(\alpha, \beta)}. \tag{4}$$

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

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

B.2 Cladistic Similarity

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 equal distance from the proto-language at Level 0. Hence, the dashed lines depict the assumed relationship between the proto-language (Level 0) and all current Eritrean languages (Level 6).

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:

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

where δ is an arbitrarily assigned weight used to discount more recent linguistic cleavages relative to deep cleavages. I describe this weight as arbitrary because there is no consensus on the appropriate weight to be assumed. 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 have constructed a phylogenetic language tree for the 8 distinct languages of Eritrea. The dashed lines represent this assumed historical relationship, so in all cases the contemporary Eritrean languages possess an equal number of branches to the proto-language at Level 0. Although $M = 6$ in Figure B1, in the Ethnologue language tree the highest number of classifications for any language is $M = 15$, which I abstract from here for simplicity.

B.3 Coethnicity

Coethnicity is a dummy variable that is equal to one when a partitioned group is the same ethnolinguistic group that a leader descends from, i.e., linguistic similarity is equal to one.

$$\text{Coethnicity} = \begin{cases} 1 & \text{if linguistic similarity} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

C Mapping Ethnicity to Language

There is mostly agreement between ethnographers that language is a suitable marker of ethnicity in Africa (Batibo, 2005; Desmet et al., 2017). The challenge of mapping ethnicity to language is that, in some instances, a single ethnic group speaks many languages. In such instances it's not obvious what language is the appropriate language to match to a leader's ethnicity. As a solution to this problem I use the following three-step assignment rule to construct a mapping between ethnicity and language in Africa.

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

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

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

D Supplementary Material

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

D.1 Various Fixed Effects Specifications

Table D1 reports 27 different estimates: 9 versions of equation (1) from the paper for each of the 3 linguistic similarity measures. Columns 1-3 report between-group estimates with country-year fixed effects, the estimates in columns 4-6 add country-language fixed effects, and columns 7-9 report estimates for the triple-difference estimator. For each set of three regressions I report estimates (i) without any covariates, (ii) estimates that only control for log population density and the logged geodesic distance between each partitioned group and the corresponding leader's group, and (iii) the full set of covariates I outlined in Section II.

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

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

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

D.2 Heterogeneity

The analysis reveals little evidence of heterogeneity (Table D2). One explanation for a lack of heterogeneity is that these different channels are only relevant in some countries and do not generalize to the 35-country sample I use here. Another possible explanation is that the rich set of fixed effects in each regression absorb much of the important variation. For

example, in column (1), I find that democracy has a mitigating effect on the extent of observed favoritism, but this effect is not statistically significant. While the intuition is consistent with Burgess et al. (2015), the lack of precision likely comes from the fact that country-year fixed effects account for the level effect of democracy, and the residual variation is not significant enough to identify any meaningful effect. A similar explanation applies to the remaining variables, where country-language fixed effects absorb the level effect for each because of the time invariance of these group-level measures.

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

D.3 Additional Controls

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

Table D3 reports these subsample estimates that include the additional control variables. For each of the three measures of similarity I report estimates that include the absolute difference in the Malaria Ecology Index, the absolute difference in the Agricultural Suitability Index and estimates that include both measures, in addition to benchmark set of controls. The results are unchanged by including these controls.

D.4 Sample Selection

My inability to observe the lexicostatistical similarity of the 64 language groups without an ASJP language list raises the question whether these unobserved groups are systematically different than those in my benchmark sample. To address this concern I test for mean differences in key observables and report these differences in Table D4.

First I show that there is no difference in the average night light luminosity between

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

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

D.5 Measurement Error

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

In this section I report estimates on a subsample of my benchmark dataset that excludes the 4 leaders I could not unambiguously match.[4] Table D5 reports these results. Overall little is changed from my benchmark estimates, with the exception that coethnicity is no longer significant at standard levels of confidence. However, lexicostatistical similarity is roust to these excluded leaders, and most importantly, column (4) of Table D5 makes clear that the significance of non-coethnic similarity is not a consequence of the possible measurement error introduced when assigning an ethnolinguistic identity to the aforementioned leaders.

D.6 Balanced Panel

In this section I test the robustness of the benchmark estimates using a balanced panel of country-language groups between 1992 and 2013. My benchmark panel was unbalanced because of missing data on language lists used to estimate lexciostatistical similarity. This is problematic if these lists are missing for non-random reasons (Cameron and Trivedi, 2005). To check this I limit the analysis to a balanced sample of 84 language groups partitioned across 23 countries. Table D6 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

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

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.

D.7 Weighted Regressions

In this section I test for heteroskedasticity in my benchmark estimates by weighting regressions by the Ethnologue population of each language group. The idea is that the measure of night light intensity is an average within each country-language group, and it is likely to have more variance in places where the population is small (Solon et al., 2015). Table D7 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.

D.8 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 choice to add 0.01 to the log transformation. Nonetheless, I experiment with two alternative transformations in Table D8.

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.

D.9 First Differences

I report first difference estimates in Table D9. While I do find a positive coefficient for each measure of similarity, the majority of estimates fall just outside standard levels of confidence. This is due to the fact that there is less variation in changes of similarity over time than there is across groups in levels.

D.10 DHS Additional Tables

Table D12 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.[5] The magnitude of locational similarity increases slightly and maintains its strong significant effect on individual wealth. In Table D12 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 the geodesic distance to the nearest coast and border.[6]

The middle panel of Table D10 reports estimates for individual similarity. While all coefficients take the expected positive sign, only a single estimate of individual similarity is statistically significant. When I do not control for any covariates the effect of individual similarity is very precisely estimated. To the contrary, the effect goes away once I account for respondents living in rural locations. The same is true when including the full set of controls.

Next I jointly estimate both channels using the aforementioned variation among individuals non-native to the region in which they reside. The results are consistent with the rest of the table and reported in the bottom panel of Table D10. In column (1) the estimate for

[5]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.

[6]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.

individual similarity outperforms locational similarity when no individual characteristics are accounted for, however the reverse is true in columns (2)-(5) as covariates are incrementally added – in particular the rural indicator.

To show that the locational mechanism is not only driven by the coethnic effect, I separately estimate locational coethnicity and non-coethnic locational similarity. I do this in the same way I did in the regional-level analysis: I define non-coethnic locational similarity as $(1 - \text{coethnicity}) \times \text{locational similarity}$. Table D11 reports these estimates. While non-coethnic locational similarity is estimated to be no different than zero in the most basic regression, once again after the baseline set of controls are added both the coethnic and non-coethnic effect are positive and strongly significant. Using the more conservative estimates of column (5), this suggests that the average level of non-coethnic locational similarity (0.164) yields an increase of 0.094 ($= 0.164 \times 0.573$) in the wealth index – roughly one fourth the coethnic effect.

Finally, I also report the DHS estimates for locational similarity and include each baseline covariate one at a time. The idea here is to highlight the relative importance of controlling for the urban-rural inequality gap when using the DHS wealth index (Young, 2013). Table D12 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.

D.11 Coalition Building

Data

I use data from Francois et al. (2015) on the share of an ethnic group's representation in the governing coalition for 15 African countries.[7] These data are available at a yearly interval until 2004 for the ethnic groups listed in Alesina et al. (2003) and Fearon (2003). Because the unit of observation is an ethnic group, I assign an Ethnologue language group to each ethnicity using the assignment strategy outlined in Appendix C.[8] I measure the lexicostatistical similarity of these groups to the ethnolinguistic identity of the national

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

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

leader between 1992 and 2004 using the leader data described in Section I of the paper. In each country a residual ethnic categorization named Other is assigned to capture all groups outside of a country's major ethnic groups. Because Others lack a single ethnolinguistic identity, I assign Other groups a value of zero percent similarity to their leader.

Results

I report estimates of equation (3) from the paper in Table D13. Column 1 replicates the main estimate of Francois et al. (2015) on the subset of data that I observe lexicostatistical similarity. The coefficient for coethnicity takes the expected positive sign, implying there is a 9 percent increase in the leader's group share of the governing coalition over and above the ministerial appointments made in accordance with the leader's group size. The magnitude of this coefficient is slightly smaller than the comparable coefficient in Francois et al.'s (2015) Table III. This suggests that, if anything, this subsample biases the coefficient downward. Column 2 corroborates this result using lexicostatistical similarity in place of coethnicity. In column 3, I separate the effect of coethnicity from lexicostatistical similarity using the same approach I used in Section III; i.e., non-coethnic similarity = (1 − coethnicity) × lexicostatistical similarity. The reported estimates in column 3 confirm that linguistic similarity predicts a group's representation in the governing coalition even among non-coethnic groups.

In columns 4-6 I explore the allocation of top positions in the governing coalition, and in columns 7-9 the allocation of positions outside of the top.[9] In all cases the variables of interest are positive and statistically significant. The most notable observation in this table is remarkable consistency in the magnitude of non-coethnic similarity across specifications. Related groups outside of the leader's ethnic group benefit from receiving positions both low and high in the hierarchy of government.[10]

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

[10]Though not reported here, the estimates for group size are statistically significant in all instances. The estimates are also comparable in magnitude to those in Table 3 of Francois et al. (2015), and similarity show evidence of concavity in the effect of group size.

Table D1: Benchmark Regressions Using Various Combinations of Fixed Effects

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

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

Table D2: Benchmark Regressions with Heterogeneous Effects

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

This table reports a series of tests for heterogeneous effects in the benchmark estimates. Average night light intensity is measured in language group l of country c in year t, and lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the national leader and is measured on the unit interval. All control variables are described in Table 3 of the paper. Democracy is the polity2 score of democracy for the country in which a group resides, geodesic distances are measured in kilometres from a group's centroid to the capital city and the nearest coast, oil reserve and diamond mine represent indicators variables at the group level. Standard errors are clustered at the country-language group level and are reported in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table D3: Robustness Check: Benchmark Regressions with Additional Control Variables

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

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

Table D4: Selection into Lexicostatistical Language Lists

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

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

Table D5: Robustness Check: Excluding Leaders with Ambiguous Ethnolinguistic Identities

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

This table reports estimates from a subsample that excludes all ambiguous leadership assignments. Because these problematic assignments introduce measurement error, excluding them from the analysis ensures that the results are not a consequence of measurement. Average night light intensity is measured in language group l of country c in year t, and Lexicostatistical similarity is a continuous measure of language group l's phonological similarity to the national leader and is measured on the unit interval. The same log transformation of the dependent variable is used for the lagged value of night lights, i.e., $\ln(0.01 + NightLights_{c,l,t-1})$. All control variables are described in Table 3 of the paper. 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 D6: Robustness Check: Benchmark Regressions on a Balanced Panel

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

This table reproduces benchmark estimates on a balanced subset of the panel dataset. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. All control variables are described in Table 3 of the paper. 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 D7: Robustness Check: Benchmark Regressions Weighted by Language Group Population

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

This table reports the benchmark estimates weighted by Ethnologue language group population. Average night light luminosity is measured in language group l of country c in year t, and linguistic similarity measures the similarity between each language group and the ethnolinguistic identity of country c's leader in year $t-1$. Lexicostatistical similarity is a continuous measure of a language pair's phonological similarity and is measured on the unit interval. Cladistic similarity is a discrete measure of similarity representing a language pair's ratio of shared branches on the Ethnologue language tree. Coethnic is binary measure of linguistic similarity that takes a value of 1 when language group l is also the ethnolinguistic identity of country c's leader. All control variables are described in Table 3 of the paper. 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 D8: Robustness Check: Benchmark Regressions with Alternative Dependent Variables

	$\sqrt{\text{nightLights}_{c,l,t}}$			$\ln(\text{nightLights}_{c,l,t})$			$\text{nightLights}_{c,l,t}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Lexicostatistical similarity$_{t-1}$	0.038**			0.396**			0.257		
	(0.018)			(0.191)			(0.230)		
Cladistic similarity$_{t-1}$		0.029*			0.189			0.236	
		(0.016)			(0.163)			(0.218)	
Coethnic$_{t-1}$			0.012			0.258*			0.110
			(0.014)			(0.138)			(0.162)
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-language fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	214	214	214	297	297	297
Countries	35	35	35	33	33	33	35	35	35
Language groups	164	164	164	98	98	98	153	153	153
Adjusted R^2	0.952	0.952	0.952	0.935	0.935	0.935	0.998	0.998	0.998
Observations	6,610	6,610	6,610	2,921	2,921	2,921	5,098	5,098	5,098

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 ($\ln(\text{nightLights}_{c,l,t})$). Columns (7)-(9) are estimated using a Poisson pseudo maximum likelihood estimator. 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 of the paper. 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 D9: Robustness Check: Benchmark Regressions in First Differences

Dependent Variable: ΔnightLights$_{c,l,t}$						
	(1)	(2)	(3)	(4)	(5)	(6)
ΔLexicostatistical similarity$_{t-1}$	0.018	0.018				
	(0.013)	(0.013)				
ΔCladistic similarity$_{t-1}$			0.006	0.006		
			(0.010)	(0.010)		
ΔCoethnic$_{t-1}$					0.020	0.020
					(0.014)	(0.014)
Distance & population density	No	Yes	No	Yes	No	Yes
Geographic controls	No	Yes	No	Yes	No	Yes
Language-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Country-year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Clusters	355	355	355	355	355	355
Adjusted R^2	0.330	0.329	0.330	0.329	0.330	0.329
Observations	6,255	6,255	6,255	6,255	6,255	6,255

This table reproduces benchmark estimates in first differences on the raw night lights data. 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 first differenced and described in Table 3 of the paper. 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 D10: Individual-Level Regressions: Locational and Individual Similarity

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

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

Table D11: Individual-Level Regressions: Locational and Individual Similarity

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

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

Table D12: 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 Table 8 in the paper. The unit of observation is an individual. Standard errors are in parentheses and adjusted for clustering at the country-language-wave level. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table D13: Ethnic Favoritism and Coalition Power Sharing

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

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

References

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

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.

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

Desmet, K., Ortuño-ortín, I., and Wacziarg, R. (2017). Culture, ethnicity and diversity. *American Economic Review*, 107(9).

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

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

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

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

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

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.

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.

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

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

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

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

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

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

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