Online Appendix to

Linking Ethnicity in Africa: Data and Methods

Table of Contents

A Coding Procedure

The language-based link between any two ethnic group datasets requires that each ethnic category in the two lists (Table I; main text) are mapped to the language(s) and language families associated with the group. We link about 8'100 distinct ethnic categories[1] to the tree of African languages comprising about 15'200 nodes, 2154 primary languages (level 15), and 4822 dialects (level 16). To reduce the potential for errors, we implement a structured matching procedure, double-coding each link independently and correcting inconsistencies in a third coding round. On a country-by-country basis, coders take the following steps:

Table A1: Ethnic groups from DHS in Nigeria: Excerpt

Group	Share	Match: direct	Match: alt. name	Match: dialect	Match: foreign	Match: previous
Abua	<.01	Abua [org]				
Adra/Adarawa	<.01	Adamawa [L6]		Adarawa [dial]	Adamawa [L6]	
Adun	<.01			Adun [dial]		
Afemai	<.01		Yekhee [org]			
Afizire	<.01		Izere [org]			

Notes: Column 'Match: previous' is automatically updated as matching proceeds.

1. The coder finds a table similar to Table A1 that lists all ethnic labels contained in a particular list and country, here the DHS from Nigeria. The table includes a set of automatically generated matches between the name of the group and four types of language labels.[2] All of these automatic matches are generated via fuzzy string matching,[3] and represent suggestions of decreasing quality. As Table A1 shows, the proposed direct match between the Abua group and the corresponding Ethnologue language has no rivalling suggestion. It is very likely that the Abua indeed speak Abua. In contrast, the Adra/Adarawa may by linked with the Adamawa language family or the Adarawa dialect. It takes some additional research to find the appropriate link here. Similarly, coders needed to consult additional sources to confirm whether the Afenmai do indeed speak Yekhee.

[1] This number does not include categories from the SIDE data, which are contained in the DHS data.

[2] First, we directly match names to the name of nodes on the language tree in the same country. Second, we match names to alternative names of the countries' languages. Third, we match to dialects associated with these languages. Fourth, we match the group names to these three types of language names, but now across all African countries other than the country the coder is working on.

[3] Fuzzy string matches are based on a maximum Levenshtein distance of .8.

2. Starting from the the automatic suggestions, coders establish the most appropriate link between a given ethnic category and one or more Ethnologue nodes. Coders draw on qualitative information on ethnic groups to double-check suggestions, adjudicate between contradictory automatic matches, and find matches for groups without a suggested match. Some of this information comes from the datasets themselves, such as the size of the group (Column 2 in Table A1), or descriptions of the groups in the respective codebooks.[4] Other information comes from encyclopediae such as *The Peoples of Africa: An Ethnohistorical Dictionary* (Olson, 1996). Lastly, standard online sources on ethnic groups such as Wikipedia, the Encyclopedia Britannica, and the Joshua Project are consulted as well. Table A5 below summarizes the degree to which our coders followed or deviated from automated suggestions across all data sets. If no match is found or a category refers to a non-ethnic cleavage (for example a geographic unit, a village, or even a surname) coders supply this information in a comment. Table A6 lists all unique ethnic categories for which we were unable to establish a link to the language tree.

3. As the matching of groups to languages proceeds, algorithms ensure that matched languages actually exist in Ethnologue. Additionally, each completed match is automatically transferred as a suggestion to ethnic categories with a similar name in other lists of the same country (see column 'Match: previous' in Table A1. This avoids redundant effort and increases the consistency of our coding across different datasets.

4. After all ethnic categories from all countries are linked to Ethnologue, we run a number of post-coding checks. These identify groups without a match and comment, potential inconsistencies in matchings of groups that share the same name, as well as inconsistent matchings of groups that cross borders.[5] The respective coding decisions are then double-checked and corrected if necessary.

In order to identify errors in our coding and increase its reliability, two coders follow steps 1-4 independently of each other. Cases with conflicting coding decisions are revised in a third round in which we assess the respective coders'

[4]EPR, Murdock, and in some cases AMAR offer textual descriptions of the ethnic groups and subgroups contained in the respective dataset.

[5]This last check applies only to the GREG and Murdock data. Both datasets provide maps of ethnic homelands without nesting them inside countries.

justification of their links and consult additional sources to arrive at the most appropriate link. All ethnic datasets were thus independently linked to Ethnologue twice. The only exception is Posner's (2004) PREG dataset which we added later in the process and only coded once.

B Reliability

Table A2 presents the intercoder-reliability metrics between the two initial coding rounds. We note that 70% of all coding decisions are exactly the same across coders. In 20% of all cases, coders link an ethnic category to overlapping sets of nodes in the linguistic tree. Many of these cases are caused by uncertainty about the boundaries of an ethnic category in a list and occur if, in the example in Figure 2a, coder 1 links the Akan from Afrobarometer to the Akan on level 9, while coder 2 links them to the Akan on the language level (level 15). This type of inconsistency occurs much more frequently in lists of highly aggregate ethnic groups such as EPR and Murdock, where ethnic groups are usually linked to multiple languages. In about 4% of all cases, one of the coders does not find a language while the other one does. 5% of all ethnic categories are matched to completely different linguistic nodes. This is a particular problem of the AMAR dataset, which contains many highly disaggregated ethnic categories that are described in historical dictionaries and are hard to identify on the language tree.

Table A2: Intercoder reliability: By list type

Type	N	Equal	Partial overlap	Missing link	Disjoint
All	7,991	0.70	0.20	0.04	0.05
Afrobarometer	1,582	0.78	0.13	0.05	0.05
AMAR	1,560	0.71	0.17	0.04	0.08
DHS/SIDE	1,471	0.76	0.14	0.06	0.04
EPR	298	0.59	0.31	0.08	0.03
Fearon	361	0.71	0.22	0.04	0.04
FRT	279	0.68	0.29	0.01	0.03
GREG	491	0.70	0.26	0.002	0.04
IPUMS	639	0.78	0.11	0.08	0.03
Murdock Map	1,310	0.54	0.38	0.02	0.07

How reliable is our coding with respect to existing links between ethnicity datasets? We compare our data to five existing and independent matches between different datasets and find a high degree of correspondence. The five existing matching tables consist of two unpublished links between the EPR dataset to

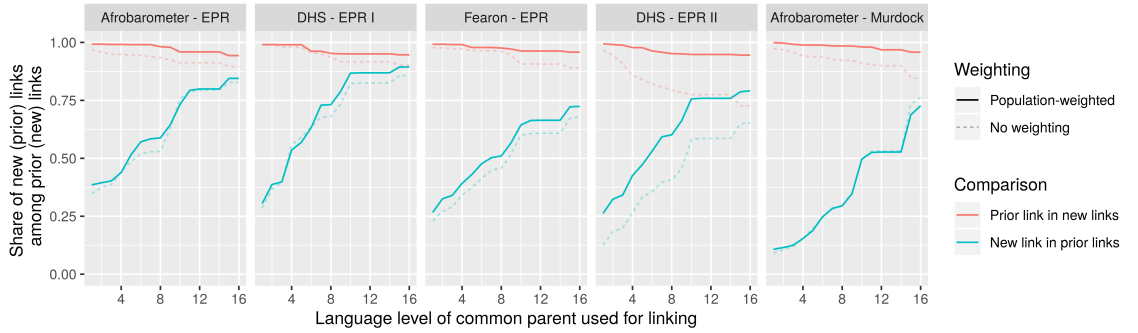

Figure A1: Recovery of previously coded links between groups by matching groups via common parent nodes at varying Ethnologue language levels (see Figure 2)

the Afrobarometer and DHS surveys, one link between EPR and Fearon's list (Cederman, Weidmann & Bormann, 2015), one link between EPR and DHS (Müller-Crepon & Hunziker, 2018), and a final link between Murdock's Map and the Afrobarometer (Nunn & Wantchekon, 2011).[6] Figure A1 plots the matches that existing efforts recover in our dataset (red) and the matches that our data collection recovers in previous efforts (blue) along the Ethnologue language tree levels from low (on the left) to high (on the right).[7]

We recover matches in existing link files in at least 90% of all cases at the highest resolution, i.e., the dialect level.[8] In contrast, prior efforts to match two distinct ethnic group lists recover our coding only to a lesser extent: at the highest linguistic resolution, we find recovery rates between a low of 72% and a maximum of 90%. The divergence is due to our language-based dictionary approach that places no restrictions on the size of required overlap between groups a and b. This yields many more one-to-many matches than encoded in previous match files.

[6]To present consistent results, we drop matches from Nunn & Wantchekon (2011) that link Afrobarometer respondents with Murdock groups outside of their country.

[7]It is easier to agree on a link if the Ethnologue resolution is low and the resulting categories correspondingly broad.

[8]Decreasing the resolution or moving up the language tree automatically increases the recovery rate as groups are matched at increasingly broad ethnic categories.

C Additional Figures and Tables

Table A3: Matched ethnic group lists

List	Inclusion Criterion	Contents
Afrobarometer	none	political, economic & social attitudes, conditions & behavior
AMAR	social relevance & population threshold	political, social, economic status; external support; conflict behavior
DHS	none	demographics, health, nutrition, economic well-being
EPR	political relevance	political representation, regional autonomy, conflict behavior
Fearon	population threshold	population shares & country-level diversity
FRT	similar but not equivalent to Fearon	ethnicity of ministers
GREG	unknown but mainly linguistic groups	settlement areas & population shares
IPUMS	official recognition by the state	demographics, education, etc.
Murdock Map	unknown	settlement areas and ethnographic variables (via Ethn. Atlas)
PREG	political relevance	population shares & country-level diversity
SIDE	based on DHS & population threshold	local-level population shares
WLMS	based on Ethnologue	settlement areas

Table A4: Linkage rates by datset and country (in percent, population weighted)

Country	Afrobarometer	AMAR	DHS	EPR	Fearon	FRT	GREG	IPUMS	Murdock Map	PREG	SIDE
BDI	96	94		94	96		93		100	94	
BEN	92	87	91	93	79	88	92		94	77	91
BFA	92	85	92	89	91		87	92	96	75	92
BWA	91	86		93	93		89		88	80	
CIV	90	79	88	83	79	86	73		89	49	80
CMR	82	78	92	85	89	87	81		89	78	93
CPV	78			93			93				
DZA	100	100		100	100		100		100		
EGY	94	83		94	94		97		98		
GAB	82	82	92	78	89	91	96		82	71	93
GHA	94	83	91	88	82	89	77	90	95	80	92
GIN	86	87	91	85	80	91	92	26	100	71	91
KEN	93	78	93	92	92	92	88		94	70	92
LBR	88	85	86	46	87	87	41	87	75	89	87
LSO	91	68		87	89		87		93	87	
MAR	100	100		100	100		100	100	100		
MDG	99	95		96	97		100		97	100	
MLI	96	92	96	99	97		90	96	97	85	92
MOZ	86	73	84	71	79		83		94	67	84
MUS	81	83		88	85		74			60	
MWI	95	92	92	92	91		85	92	93	82	92
NAM	95	89	94	95	95		91		89	87	95
NER	97	80	97	94	94		95		97	92	94
NGA	91	72	87	82	86	84	84	49	91	78	87
SDN	69	73		84	78		84		87	71	
SEN	96	86	94	92	94		94	96	96	70	95
SLE	92	84	94	73	91	91	63	90	87	73	88
STP	52										
SWZ	85	82		88	92		95		90		
TGO	88	82	89	78	86	86	87		97	82	90
TUN	93	92		93	100		99		100		
TZA	82	65		47	72	78	70		81	64	
UGA	89	77	93	77	89	89	78	91	98	68	90
ZAF	94	94	98	99	95		93	94	78	76	
ZMB	93	81	93	86	87		85	93	91	82	90
ZWE	93	79	83	79	88		92		91	68	

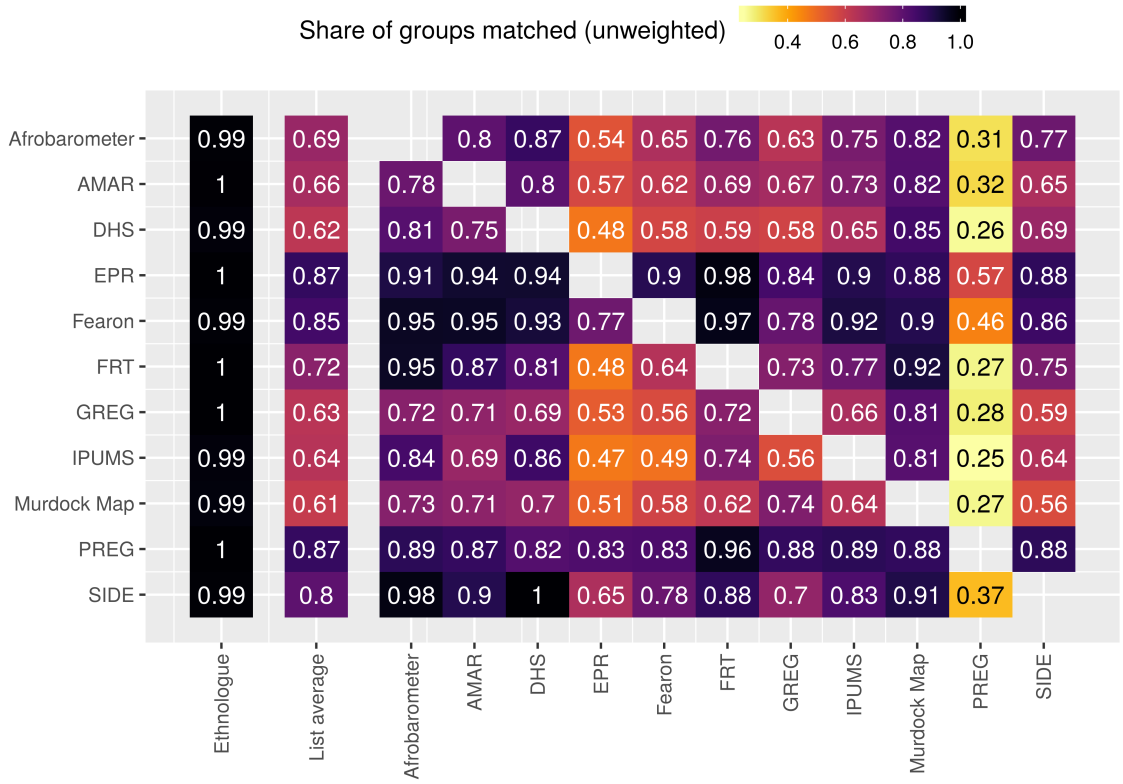

Figure A2: Proportion of groups per list matched to Ethnologue and other lists. Each ethnic category receives the same weight.

Table A5: Overlap between coded and automatically proposed matches

Type	Matches coded			Matches proposed		
	Total	same as proposed match (in %)		Total	same as coded match (in %)	
		Language name	Any name		Language name	Any name
Afrobarometer	1560	25	50	3724	83	21
AMAR	1623	28	51	4896	77	17
DHS	1326	28	52	4267	74	16
EPR	510	23	34	1305	67	13
Fearon	511	24	38	1517	65	13
FRT	372	34	47	1122	71	16
GREG	717	18	35	2479	61	10
IPUMS	534	20	39	1386	79	15
Murdock Map	1984	14	25	4833	54	10
SIDE	484	37	59	1774	75	16
Total	9621	23	42	27303	71	15

Note: 'Org. name' refers to automatically proposed matches on the basis of the names of Ethnologue's languages and the clusters they belong to. 'Any' refers to any type of automtically proposed match. Thus, in the case of Afrobarometer, of 1560 matches, 25% have been proposed automatically based on the name of an Ethnologue langauge. 50% have been proposed based on any name, alternative name or subdialect of a language, or langauge from other countries. Reversely, of the 3724 proposals made for Afrobarometer matches, only 21% have been coded as actual match.

Table A6: Groups without a match in Ethnologue

Country	Groups
AGO	Cacondas [AMAR]; Chicumas [AMAR]; Haco [AMAR]; Hungo [AMAR]; K'bala [AMAR]; Kakondas [AMAR]; Kalukembes [AMAR]; KOROCA [Murdock Map]; Luango [AMAR]; Mbondo [AMAR]
BFA	Ghanian [DHS]; Kibsi [AMAR]; Malian [DHS]; Nsp [DHS]; Pays Cedeao [DHS]
BWA	Mokgothu [Afrobarometer]; Sekgothu [Afrobarometer]
CAF	Besom [AMAR]
CIV	Apatride [DHS]; Cameroun [DHS]; Eda [AMAR]; French [Afrobarometer]; Guinee [DHS]; Guinee [SIDE]; Ivoiriens Sans Precision [DHS]; Ivoiriens Sans Precision [SIDE]; Lebanese [FRT]; Liban [DHS]; Mauritani [DHS]; Naturalise Ivoirien [DHS]
CMR	Camerounian [DHS]; Camerounian [SIDE]; Mobakoh [Afrobarometer]; Yabassi [Afrobarometer]
COD	Bas-Kasai and Kwilu-Kwngo [DHS]; Bas-Kasai and Kwilu-Kwngo [SIDE]; Bas-Kasai et Kwilu-Kwngo [DHS]; Bas-Kasai et Kwilu-Kwngo [SIDE]; Basele-k , Man. and Kivu [DHS]; Basele-k , Man. and Kivu [SIDE]; Basele-k , Man. et Kivu [DHS]; Basele-k , Man. et Kivu [SIDE]; Cuvette Central [DHS]; Cuvette Central [SIDE]; Kasai, Katanga, Tanganika [DHS]; Kasai, Katanga, Tanganika [SIDE]; Kivu Province [Fearon]; Kwilu Region [Fearon]; Ubangi and Itimbiri [DHS]; Ubangi and Itimbiri [SIDE]; Ubangi et Itimbiri [DHS]; Ubangi et Itimbiri [SIDE]; Uele Lac Albert [DHS]; Uele Lac Albert [SIDE]; Uele Lake Albert [DHS]; Uele Lake Albert [SIDE]
COG	Bahumbu [DHS]; Bakaya [DHS]; Bweni [DHS]; Europe et Oceanie [DHS]; IKASA [Murdock Map]; Kabinda [DHS]; Mayanga [DHS]; Minkengue [DHS]
CPV	Relacionado com o estado de espirito [Afrobarometer]
ETH	Djebutians [DHS]; From Different Parents [DHS]; Guagu [DHS]; Guagugna [IPUMS]; Koma / Komo, Hayahaya, Medin, Akuwma [DHS]; Wergigna [IPUMS]; Zlmamigna [IPUMS]
GHA	Brefo/Birfuo [Afrobarometer]; Feras [AMAR]; Nabi [Afrobarometer]; Nandom [Afrobarometer]; Nsahas [Afrobarometer]; Zabagle [Afrobarometer]
GIN	Manian [Afrobarometer]
KEN	Gabawen [Afrobarometer]; Garmug [Afrobarometer]; Ombuya [Afrobarometer]
LBR	No tribal affiliation [IPUMS]; None [DHS]
LSO	Balafe [Afrobarometer]; Baropoli [Afrobarometer]; Bavudie [Afrobarometer]; Ledozeni [Afrobarometer]; Lepele [Afrobarometer]; Mantsosa [Afrobarometer]; Mapele [Afrobarometer]; Mapokwana [Afrobarometer]; Mbokwakoana [Afrobarometer]; Mchegu [Afrobarometer]; Mochrist (Jesus) [Afrobarometer]; Mokhalo [Afrobarometer]; Mokhatla [Afrobarometer]; Mokhebesi [Afrobarometer]; Monareng [Afrobarometer]; Mopeli [Afrobarometer]; Mophiring [Afrobarometer]; Motaung [Afrobarometer]; Motebang [Afrobarometer]; Motsoeneng [Afrobarometer]; Mzema [Afrobarometer]; Sephotsa [Afrobarometer]
MDG	langue regionale [Afrobarometer]; Tealaotra [Afrobarometer]; Zaza lava mahafasa [Afrobarometer]
MLI	Cdeao Country [DHS]; Ecowas Countries [DHS]; Ecowas Countries [SIDE]; Ne Sait Pas [DHS]; Non Malian [DHS]; Trouka [Afrobarometer]
MOZ	Islamic Coastal [Fearon]; Zambezi [Fearon]
MUS	Muslims [EPR]

NGA	Agazawa [DHS]; Ahu [DHS]; Amamong [DHS]; Awo [DHS]; Bafeke [DHS]; Bagathiya [DHS]; Bageri [DHS]; Bagunge/Badagire [DHS]; Bahnake [DHS]; Baji/Biji [DHS]; Barabaci [DHS]; Bayam [Afrobarometer]; Beteer [DHS]; Buko [DHS]; Chiba [DHS]; Dumak [DHS]; Eterco [Afrobarometer]; Etina [DHS]; Foron [DHS]; Gmenchi [DHS]; Gomo/Gamoyaya [DHS]; Gumbarawa [DHS]; Gwoza [Afrobarometer]; Gwoza [DHS]; Hanbagda [DHS]; Igbanko [DHS]; Ijeme [DHS]; Ikara [DHS]; Jajiri [DHS]; Kantanawa [DHS]; Knale [Afrobarometer]; Kuba [Afrobarometer]; Kunkawa/Kawa [DHS]; Mangus/Manju [DHS]; Mbwa [DHS]; Mgas [Afrobarometer]; Mirnang [DHS]; Muryan [DHS]; Nanba/Wanba [Afrobarometer]; Nezou [DHS]; Nkwana [Afrobarometer]; Nnebe [DHS]; Normana [Afrobarometer]; Obubua [DHS]; Odu [DHS]; Ogbo [DHS]; Ohari [DHS]; Omele [DHS]; Paibun [DHS]; Pasama [DHS]; Rulere [DHS]; Sekere [DHS]; Somunka [DHS]; Taira [DHS]; Tangoa [Afrobarometer]; Uhionigbe [DHS]; Uru [DHS]; Uyo [DHS]; Yendre [DHS]; Yonubi [DHS]
SLE	None [IPUMS]
TCD	Fitri-Batha [DHS]; Kanem-Bornou [DHS]; Kebbi [DHS]; Lac Iro [DHS]; Mayo Kebbi [DHS]; Tandjile [DHS]
TGO	Aklobo [Afrobarometer]; Ndebele [Afrobarometer]; Stranger [DHS]; Stranger [SIDE]
UGA	Aliba [Afrobarometer]; Aliba [DHS]; Bakonki [DHS]; Banahaabi-Hayo [DHS]; Batoro, Batuku, Basongora [IPUMS]; Birugi-Muyinda-Mwega [DHS]; Bowa-Muwaya [DHS]; Digging [DHS]; Goanese [AMAR]; Middle East [IPUMS]; Mulalo [DHS]; Ngirivu-Gisi [DHS]; Oceania [IPUMS]; Reli [DHS]
ZAF	Asian [Fearon]; Asians [EPR]; Shangaan/Tsonga/Ronga/Tswa [Afrobarometer]
ZMB	American [DHS]; American [IPUMS]; Asian [DHS]; Asian [IPUMS]; Asian language [IPUMS]; European [DHS]; European [IPUMS]; European language [IPUMS]; North-Western [DHS]
ZWE	Asian [DHS]; Vhitori [Afrobarometer]

Table A7: Mistrust in President: EPR & FRT

	Mistrust in President					
	EPR			FRT		
	(1)	(2)	(3)	(4)	(5)	(6)
Ethnic Link to Gov.	-0.359^{***}		-0.226			
	(0.096)		(0.149)			
Ling. Dist. to Gov.		0.400^{*}	0.221			
		(0.165)	(0.250)			
Ethnic Link to Leader				-0.275^{**}		-0.080
				(0.099)		(0.124)
Ling. Dist. to Leader					0.392^{*}	0.342
					(0.156)	(0.186)
Country-Survey FE	yes	yes	yes	yes	yes	yes
Ethnic Group FE	no	no	no	no	no	no
Observations	8,653	8,653	8,653	8,653	8,653	8,653
Adjusted R^2	0.314	0.312	0.318	0.299	0.309	0.310

Notes: Dependent variable standardized to mean 0 and sd 1. Control variables include age, age squared, education level indicators, a female and an urban dummy. Standard errors clustered on ethnic group in parentheses. Significance codes: *p<0.05; **p<0.01; ***p<0.001

Table A8: Ethnic Grievances: EPR & FRT

	Unfair treatment of own group					
	EPR			FRT		
	(1)	(2)	(3)	(4)	(5)	(6)
Ethnic Link to Gov.	-0.369^{***}		-0.269			
	(0.079)		(0.161)			
Ling. Dist. to Gov.		0.375^{*}	0.166			
		(0.150)	(0.257)			
Ethnic Link to Leader				-0.288^{**}		-0.123
				(0.109)		(0.143)
Ling. Dist. to Leader					0.364^{*}	0.286
					(0.152)	(0.196)
Country-Survey FE	yes	yes	yes	yes	yes	yes
Ethnic Group FE	no	no	no	no	no	no
Observations	7,148	7,148	7,148	7,148	7,148	7,148
Adjusted R^2	0.104	0.099	0.105	0.092	0.097	0.098

Notes: Dependent variable standardized to mean 0 and sd 1. Control variables include age, age squared, education level indicators, a female and an urban dummy. Standard errors clustered on ethnic group in parentheses. Significance codes: *p<0.05; **p<0.01; ***p<0.001

LEDA R-Package Documentation

Initialize linking object

The LEDA package is programmed in an object oriented manner. Once you initialize a LEDA-object, methods are applied directly to the object and either change the object or return the results of a query. See the documentation of the R-package R6 for details.

Create LEDA objects

```
library(LEDA)
leda <- LEDA$new()
```

Help files

Because all functionalities of the LEDA package are methods of LEDA objects, all documentation can be accessed by calling `?LEDA`.

Datasets included in LEDA

To get a first overview of the possibilities coming with LEDA, start querying the 'list dictionary', which contains all metadata of all lists of ethnic groups that the LEDA project links to the Ethnologue language tree. Lists are identified by their country, the type of dataset (e.g. EPR, Afrobarometer, DHS), the variable that identifies ethnic groups in that dataset, the type of ethnic marker (language, ethnic group, mother tongue), as well as year or survey-round identifiers where appropriate.

```
# Retrieve dataset dictionary
list.dict <- leda$get_list_dict()
# Show first entries
head(list.dict)
```

```
##     list.id type cowcode iso3c       marker groupvar year round subround
## 1:1       1 AMAR     404   GNB ethnic group    Group   NA    NA       NA
## 1:2       2 AMAR     420   GMB ethnic group    Group   NA    NA       NA
## 1:3       3 AMAR     432   MLI ethnic group    Group   NA    NA       NA
## 1:4       4 AMAR     433   SEN ethnic group    Group   NA    NA       NA
## 1:5       5 AMAR     434   BEN ethnic group    Group   NA    NA       NA
## 1:6       6 AMAR     435   MRT ethnic group    Group   NA    NA       NA
```

```
# All data types
unique(list.dict$type)
```

```
##  [1] "AMAR"          "DHS"           "SIDE"          "EPR"
##  [5] "Fearon"        "FRT"           "GREG"          "Murdock_Map"
##  [9] "IPUMS"         "Afrobarometer" "WLMS"          "PREG"
```

Link data sets

Once familiar with the lists of ethnic groups that are part of the LEDA object, we can proceed to link the groups contained in any two lists of groups to each other. The LEDA object includes three methods to link

lists of ethnic groups to each other, each of them described below.

Link via set relations

We can first link lists A to lists B by analyzing the set of nodes on the language tree that groups a and b share. In the example below, we link two groups to each other as soon as they are associated with at least one common dialect on the language tree (`link.level = "dialect"`). As one specifies link levels closer to the root of the language tree, i.e. by setting `link.level = "language"` or `link.level = 5` (language tree level 5 of 16), the number of groups b linked to a increases and links become less precise.

The lists entered for parameters `lists.a` and `lists.b` offer a flexible way to select the lists of ethnic groups that are linked to each other. Note that you can enter any parameter combination that identifies at least one list of ethnic groups, but potentially many. The latter is helpful if you want to, for example, link all Afrobarometer surveys to the Ethnic Power Relations (EPR) data. It is generally (but not always) sensible to only link lists of ethnic groups within the same country borders by setting `by.country = T`.

```
## Link all Afrobarometer groups (rounds 1-5) in Uganda to the FRT data.
setlink <- leda$link_set(lists.a = list(type = c("Afrobarometer"),
                                        iso3c = c("UGA"),
                                        round = 4, marker = "language"),
                                 lists.b = list(type = c("FRT"),
                                                iso3c = c("UGA")),
                                 link.level = "dialect",
                                 by.country = T,
                                 drop.a.threshold = 0,
                                 drop.b.threshold = 0,
                                 drop.ethno.id = T)
## Have a look
head(setlink[, c("a.group", "b.group", "a.type", "b.type")])
```

```
##      a.group b.group        a.type b.type
## 1     Acholi  Acholi Afrobarometer    FRT
## 2       Alur    Alur Afrobarometer    FRT
## 3      Ateso    Teso Afrobarometer    FRT
## 4 Japhadhola Padhola Afrobarometer    FRT
## 5      Kakwa   Kakwa Afrobarometer    FRT
## 6  Kiswahili    <NA> Afrobarometer   <NA>
```

One can further refine the link by constraining the arguments `drop.a.threshold` and `drop.b.threshold` that control the shares of common languages associated with groups a and b for a link to be realized. For eaxample, setting `drop.a.threshold = .5` ensures that in each link the language nodes of group b cover more than 50 percent of the language nodes associated with a. Conversely, setting `drop.b.threshold = .5` will ensure that in each pair of linked group a and b, group a covers more than 50 percent of the language nodes of b. More complex set relations can be implemented by setting the thresholds to 0 and switching `drop.ethno.id = FALSE`. The returned link table will then have multiple rows per linked pair of groups a and b, each coming with the ID of the language node they share.

Link via linguistic distances

We can also make direct use of the language tree and link groups in lists A and B on the basis of their linguistic distances to each other. To do so, LEDA calculates linguistic distances first and then subsets the distance matrix to return the links queried by the user.

Compute linguistic distance between groups

The algorithm computes the full linguistic distance matrix between groups in lists A and B. Via the parameter `level`, users can specify whether they want links to be based on distances between ethnic groups' `"language"` or `"dialect`. As before, it is sensible to not link lists across country borders by setting `by.country = T`.

The linguistic distance between two languages or dialects L_1 and L_2 is computed as :

$1 - ((d(L_1, R) + d(L_2, R) - d(L_1, L_2))/(d(L_1, R) + d(L_2, R)))^{\delta}$

where $d(L_i, R)$ is the length of path from a language to the tree's origin and $d(L_1, L_2)$ is the length of the shortest path from the first to the second language. δ is an exponent to discount short distances on the tree, reflected in the parameter `delta` below. Lastly, there are two ways to locate languages and dialects on the language tree. In the first, languages that are immediate children of a node that is located at level 4 of the language tree remain at their original level 5 (`expand = FALSE`). In the second way, the tree is expanded, and all languages are located on level 15 and all dialects on level 16. This expansion of the tree naturally changes computed linguistc distances.

Because ethnic groups are often linked to multiple languages or dialects, there can be multiple linguistic distances between any group a and b. `agg_fun.a` and `agg_fun.b` control the aggregation of these distances. `agg_fun.a` determines for any language node in a how its distances to nodes of b are aggregated. `agg_fun.b` controls how the resulting distances between nodes in a and group b are aggregated to arrive at a single distance between a and b.

```
## Compute distances
distance.df <- leda$ling_distance(lists.a = list(type = c("Afrobarometer"),
                                                 iso3c = "UGA",
                                                 round = 4, marker = "language"),
                                  lists.b = list(type = c("FRT"), iso3c = "UGA"),
                                  level = "dialect", by.country = T,
                                  delta = .5, expand = FALSE,
                                  agg_fun.a = min, agg_fun.b = min)
## Have a look
head(distance.df[, c("a.group", "b.group", "a.type", "b.type", "distance")])
```

```
##                       a.group b.group        a.type b.type  distance
## Afrobarometer.94664    Acholi  Acholi Afrobarometer    FRT 0.0000000
## Afrobarometer.94664.1  Acholi    Alur Afrobarometer    FRT 0.1471971
## Afrobarometer.94664.2  Acholi  Ankole Afrobarometer    FRT 1.0000000
## Afrobarometer.94664.3  Acholi   Ganda Afrobarometer    FRT 1.0000000
## Afrobarometer.94664.4  Acholi    Gisu Afrobarometer    FRT 1.0000000
## Afrobarometer.94664.5  Acholi   Gwere Afrobarometer    FRT 1.0000000
```

Link to closest linguistic neighbours

Based on the linguistic distances computed as discussed above, users can query, for every group a in lists A and for every list B, the closest linguistic neighbor b. Note that more than one nearest linguistic neighbor is returned wherever two or more closest groups b have the exact same lingusitic to a.

```
mindistlink <- leda$link_minlingdist(lists.a = list(type = c("Afrobarometer"),
                                                    iso3c = "UGA",
                                                    round = 4, marker = "language"),
                                     lists.b = list(type = c("FRT"), iso3c = "UGA"),
                                     level = "dialect",
                                     by.country = T,
                                     expand = FALSE,
                                     delta = .5,
```

```
                                        agg_fun.a = min, agg_fun.b = min)
## Have a look
head(mindistlink[, c("a.group", "b.group", "a.type", "b.type", "distance")])
```

```
##      a.group b.group        a.type b.type  distance
## 1     Acholi  Acholi Afrobarometer    FRT 0.0000000
## 2       Alur    Alur Afrobarometer    FRT 0.0000000
## 3      Ateso    Teso Afrobarometer    FRT 0.0000000
## 4 Japhadhola Padhola Afrobarometer    FRT 0.0000000
## 5      Kakwa   Kakwa Afrobarometer    FRT 0.0000000
## 6  Kiswahili   Gwere Afrobarometer    FRT 0.1659423
```

Link within linguistic distance

Instead of focusing on nearest linguistic neighbors only, users can also query, for every group a in lists A and for every list B, those groups b that fall within a specified distance `max.distance` of group a.

```
withindistlink <- leda$link_withinlingdist(lists.a = list(type = c("Afrobarometer"),
                                                          iso3c = "UGA",
                                                          round = 4, marker = "language"),
                                           lists.b = list(type = c("FRT"), iso3c = "UGA"),
                                           level = "dialect", max.distance = .1,
                                           by.country = T,
                                           delta = .5, expand = FALSE,
                                           agg_fun.a = min, agg_fun.b = min)
## Have a look
head(withindistlink[, c("a.group", "b.group", "a.type", "b.type", "distance")])
```

```
##      a.group b.group        a.type b.type  distance
## 1     Acholi  Acholi Afrobarometer    FRT 0.0000000
## 2     Acholi   Lango Afrobarometer    FRT 0.0741799
## 3       Alur    Alur Afrobarometer    FRT 0.0000000
## 4      Ateso    Teso Afrobarometer    FRT 0.0000000
## 5 Japhadhola Padhola Afrobarometer    FRT 0.0000000
## 6      Kakwa   Kakwa Afrobarometer    FRT 0.0000000
```

Inspect coding of the ethnic group <–> language link

Sometimes, one might want to inspect the origins of a link between to groups. LEDA allows that by giving access to the entire raw data that underlies each match. You can query the link between any list of groups and the language tree with the following method.

The resulting table contains one column `link` that contains the language tree nodes linked to any group. Note that in cases of multiple links, they are separated by a '|'. In most cases, the level of a node on the language tree is indicated in squared brackets behind the nodes name. L1 to L14 indicate super-languages, 'lang' denotes languages, 'iso' language isocodes, and 'dial' refers to dialects.

```
## Query raw link data
raw_ethno_links <- leda$get_raw_ethnolinks(param_list = list(type = "Afrobarometer",
                                                             round = 4,
                                                             marker = "language",
                                                             iso3c = "UGA"))
## Have a look
head(raw_ethno_links[, c("type","group", "link")])
```

```
##                           type      group          link
## Afrobarometer.1 Afrobarometer     Acholi  Acholi [org]
## Afrobarometer.2 Afrobarometer       Alur     Alur [L9]
## Afrobarometer.3 Afrobarometer      Ateso     Teso [L7]
## Afrobarometer.4 Afrobarometer Japhadhola   Adhola [L7]
## Afrobarometer.5 Afrobarometer      Kakwa   Kakwa [org]
## Afrobarometer.6 Afrobarometer  Kiswahili Swahili [org]
```

Add new links from groups to language tree

Having gained familiarity with the available ethnic links and methods, users can go a step further and link new lists of ethnic groups to the language tree. Doing so allows to link the new list of ethnic groups to every other list of ethnic groups covered by LEDA or independently added before.

Prepare new links between ethnic groups and the tree

First, one has to hand-code the link between ethnic groups and the language tree. However, this may be less tedious than it sounds. Via the method `LEDA$prepare_newlink_table()` one can access automatically generated suggestions to which language node(s) a particular group may link. These suggestions are generated via a fuzzy string match of a group's name to the names of (1) language nodes themselves, and (2) the names of ethnic groups already matched to the language tree. Thus, with every additional list of ethnic groups added to the data, linking new ones to the language tree becomes easier.

Once generated as shown below, the link table should be saved and the final links between ethnic groups and language nodes established by hand. I.e., users have to fill in the column `link`, using the information from the automatically generated suggestions, as well as secondary sources.

```
## Make or load some dataset of ethnic groups
new.groups.df <- data.frame(group_name = c("Alur", "Iteso", "Kakwa"),
                            iso3c = c("UGA"),
                            marker = "ethnic group",
                            stringsAsFactors = F)
## Prepare a new link table
##   This table contains suggested links between each ethnic group
##   and language nodes. The columns "link", "comment", and "source"
##   have to be filled by hand and correspond to the final link to
##   a set of language nodes (separated by '|'), comments on the link,
##   and a source (if required).
newlink.df <- leda$prepare_newlink_table(group.df = new.groups.df,
                           groupvar = "group_name",
                           by.country = TRUE,
                           return = TRUE,
                           save.path =  NULL, overwrite = T,
                           prev_link_param_list = NULL,
                           levenshtein.threshold = .2,
                           levenshtein.costs = c(insertions = 1,deletions = 1, substitutions = 1))
newlink.df
```

```
##   group_name iso3c       marker group        auto_link_org auto_link_alt
## 1       Alur   UGA ethnic group  Alur
## 2      Iteso   UGA ethnic group Iteso Teso [org]|Teso [L7]    Teso [org]
## 3      Kakwa   UGA ethnic group Kakwa          Kakwa [org]   Kakwa [org]
##   auto_link_dial       auto_link_prev
## 1                           Alur [L9]
```

```
## 2                Teso [org]|Teso [L7]
## 3                         Kakwa [org]
##
## 1
## 2
## 3 Org: Akwa [org]|Kabwa [org]|--|Alt: Kako [org]|Kwa' [org]|Teke-Kukuya [org]|Avikam [org]|--|Dial:
##   link comment source
## 1 <NA>    <NA>   <NA>
## 2 <NA>    <NA>   <NA>
## 3 <NA>    <NA>   <NA>
```

Add new links to a LEDA object

Having hand-coded the link between the new list of ethnic groups and the language tree, one can now add the new list of groups to the LEDA object. The list now enters the object in the same manner as all 'native' LEDA lists, as well as any lists added beforehand.

```
## First we need to encode links to the lanugage tree:
newlink.df$link[newlink.df$group == "Alur"] <- "Alur [L9]"
newlink.df$link[newlink.df$group == "Iteso"] <- "Teso [L7]"
newlink.df$link[newlink.df$group == "Kakwa"] <- "Kakwa [org]"
newlink.df$comment[newlink.df$group == "Kakwa"] <- "Kakwa same language as Bari, differs between langua
## Add to LEDA
leda$add_tree_links(tree.link.df = newlink.df,
                    idvars = c("iso3c", "marker"),
                    type = "My Survey")
```

```
## [1] "Added 1 lists to list dictionary"
## [1] "Added new entries to link dictionary."
```

```
## Check type list
print(unique(leda$get_list_dict()$type))
```

```
##  [1] "AMAR"          "DHS"           "SIDE"          "EPR"
##  [5] "Fearon"        "FRT"           "GREG"          "Murdock_Map"
##  [9] "IPUMS"         "Afrobarometer" "WLMS"          "PREG"
## [13] "My Survey"
```

For full traceability, the newly coded data is now also available in the raw data attached to LEDA and can be queried accordingly:

```
## Query raw link data
raw_ethno_links <- leda$get_raw_ethnolinks(param_list = list(type = "My Survey"))
## Have a look
head(raw_ethno_links[, c("type","group", "link")])
```

```
##                  type group        link
## My Survey.1 My Survey  Alur   Alur [L9]
## My Survey.2 My Survey Iteso   Teso [L7]
## My Survey.3 My Survey Kakwa Kakwa [org]
```

Join own data with other ethnic group lists

The new list can now be linked to any other list of ethnic groups in the LEDA object, in the same way as discussed above.

```
## Get set link from my survey to FRT
setlink <- leda$link_set(lists.a = list(type = c("My Survey"), iso3c = "UGA"),
                         lists.b = list(type = c("FRT"), iso3c = "UGA"),
                         link.level = "dialect", by.country = T,
                         drop.a.threshold = 0, drop.b.threshold = 0)
## Have a look
head(setlink[, c("a.group", "b.group", "a.type", "b.type")])
```

```
##   a.group b.group    a.type b.type
## 1    Alur    Alur My Survey    FRT
## 2   Iteso    Teso My Survey    FRT
## 3   Kakwa   Kakwa My Survey    FRT
```

Submit new lists to LEDA project

Given that the value of LEDA increases exponentially with the number of lists available in the R-package, we would greatly appreciate if you could share any new lists that you link to the language tree. New lists can be new rounds of survey data (e.g. Afrobarometer, DHS) or any list of ethnic groups that is based on publicly available data. You can do so by sending us an email to author /at/ xxxxx or opening an issue with the attached link file via LEDA's Github page. Shared link files should have the format returned by the method `LEDA$prepare_newlink_table()` and have the `link` column filled wherever possible.

References

Cederman, Lars-Erik; Nils Weidmann & Nils-Christian Bormann (2015) Triangulating horizontal inequality: Toward improved conflict analysis. *Journal of Peace Research* 52(6): 806–821.

Müller-Crepon, Carl & Philipp Hunziker (2018) New spatial data on ethnicity: Introducing SIDE. *Journal of Peace Research* 55(5): 687–698.

Nunn, Nathan & Leonard Wantchekon (2011) The slave trade and the origins of mistrust in Africa. *The American Economic Review* 101(7): 3221–3252.

Olson, James Stuart (1996) *The Peoples of Africa: An Ethnohistorical Dictionary*. Greenwood Publishing Group.

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