Linking Ethnic Data from Africa *

CARL MÜLLER-CREPON, YANNICK PENGL AND NILS-CHRISTIAN BORMANN

Social scientists increasingly combine multiple datasets to study ethnicity in Africa. We facilitate these efforts by systematically linking over 8'100 ethnic categories from eleven databases including surveys, geographic data, and expert-coded lists. Exploiting the linguistic tree from the Ethnologue database, we propose a systematic solution to the grouping problem of ethnicity. Novel empirical results on trust in African heads of states highlight the importance of explicitly considering sample inclusion criteria and different ways of linking ethnic categories from multiple datasets. An R-package allows researchers to link ethnic groups from any database with explicit rules and to easily add their own data on ethnic groups.

*Carl-Müller-Crepon is a Post-Doc in the International Conflict Research group at ETH Zürich, Haldeneggsteig 4 8092 Zurich, Switzerland (carl.mueller-crepon@icr.gess.ethz.ch). Yannick Pengl is a Post-Doc in the International Conflict Research group at ETH Zürich, Haldeneggsteig 4, 8092 Zurich, Switzerland (yannick.pengl@icr.gess.ethz.ch). Nils-Christian Bormann is Senior Lecturer in the Department of Government, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. We thank Paola Galano Toro, Vanessa Kellerhals, Benjamin Füglister, Lukas Dick, Carlos Mairoce, and Julian Seitlinger for invaluable research assistance. We are furthermore grateful for comments and suggestions from Levke Aduda, Matthew Gichohi, and participants of the 2018 AFK Workshop in Hamburg and the 2019 Annual Meeting of the American Political Science Association.

Ethnic identity constitutes one of the most salient political cleavages in developing countries, in particular in Sub-Saharan Africa. Not surprisingly, social scientists investigate the effect of ethnic differences on outcomes such as national identification (Robinson 2014), trust (Nunn and Wantchekon 2011), voting (Huber 2012), distributive politics (De Luca et al. 2018), and armed conflict (Østby 2008). Combining multiple meso- and micro-level datasets enables scholars to explore the effects of ethnic group-level characteristics on individual outcomes (Franck and Rainer 2012), to measure group-level attributes through micro-data (Cederman, Weidmann, and Bormann 2015), or to enrich one meso-level dataset with information from another (Wig 2016).

Linking ethnic categories from two datasets is, however, inherently difficult, because no common definition of the universe of ethnic groups exists.[1] Thus, any social scientist faces the "grouping problem" of ethnic identities (Posner 2004a, 850-1). Put differently, each dataset comes with its own list and resolution of ethnic categories. Some, for example the Ethnic Power Relations data (EPR; Vogt et al. 2015), code high-level umbrella groups that combine multiple ethnic identities. Others, such as the All Minorities at Risk data (AMAR; Birnir et al. 2014), identify as many categories as possible. Most individual-level data such as the Demographic and Health Surveys (DHS) identify the language spoken by their respondents. Thus, ethnic categories across different datasets do not easily map onto one another.

In this research note, we introduce the Linking Ethnic Data from Africa (LEDA) project. Using a dictionary-based linking procedure, we match more than 8'100 ethnic categories from the eleven most prominent datasets on ethnic groups in Africa via the

[1]We use the terms "linking" and "matching" interchangeably to describe the process of connecting any two ethnic categories from different data sources.

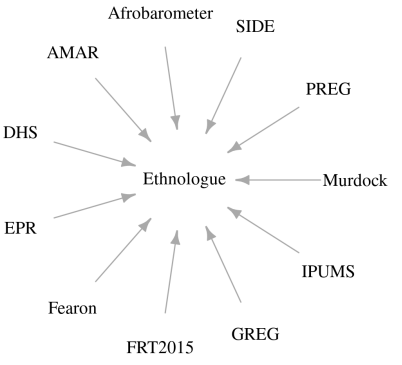


Figure 1. Meta-structure of the dictionary approach.
Data sources: Afrobarometer (2018); AMAR: Birnir et al. (2014); DHS (2018); EPR: Vogt et al. (2015); Fearon (2003); FRT: Francois, Rainer, and Trebbi (2015); GREG: Weidmann, Rød, and Cederman (2010), IPUMS: Minnesota Population Center (2017); Murdock (1959); PREG: Posner (2004a); SIDE: Müller-Crepon and Hunziker (2018).

list of known language families, languages, and dialects from the 16th edition of the Ethnologue database (Lewis 2009). The linguistic tree enables us to link groups at different resolutions, gauge the degree of linguistic overlap between any two groups, and create continuous measures of linguistic distance between them. Figure 1 depicts our approach and lists the datasets linked to each other.

Our approach to linking linguistic categories addresses three important conceptual and operational concerns about current practices. First, scholars who merge two datasets hard-code several decisions into their data, such as the resolution at which ethnic groups are linked, whether to provide many-to-many, one-to-many, or many-to-one links, and the level of overlap required for a successful match between two groups. Second, the matching tables are usually not accessible to other researchers, which limits replication attempts. As we will demonstrate below, different linking rules substantially affect empirical results. Finally, the current fragmentation of links between ethnic group lists makes it difficult to

systematically leverage the common information they contain to provide the links between existing and new ethnic group lists. Each time new ethnic links are needed, researchers start coding from scratch. Our R package LEDA[2] allows researchers to query different links between any two existing datasets, and to add new data to the language tree, thus creating links to all eleven datasets of ethnic identity that are already covered.

The grouping problem and its solution

The grouping problem of ethnic identities highlights multiple characteristics of an optimal link between two sets of ethnic groups $a \in A$ and $b \in B$. First, different datasets classify ethnic groups at different resolutions, and attempts to merge two group lists must accommodate that group a might encompass or be part of any group b. Second, ethnic categories are not necessarily fully nested within one another. Hence, the procedure must allow that a be composed of subsets of various groups in B. The optimal match is therefore many-to-many and provides information about the set relation between a group and its matches. Third, any combination of two datasets optimally goes beyond a binary logic of link or no link, and computes the distance between two ethnic categories. For example, the west-African Asante are more distant from the Yoruba than from the Fante, who, together with the Asante, belong to the ethnic cluster of the Akans.

The first step towards solving the grouping problem is to limit ourselves to linguistic identity categories. Although individuals in Africa subscribe to multiple, frequently context-dependent ethnic categories including tribe, religion, and race (Posner 2004b; McCauley 2014), language is arguably the most wide-spread identity marker globally

[2]Available at `https://github.com/carl-mc/LEDA`.

(Gellner 1983), and is particularly pronounced in Sub-Saharan Africa due to missionary activity (Vail 1989). More importantly, other ethnic markers often closely align with linguistic borders on the continent.[3] In many cases language mirrors tribal affiliations at the local level, and thus yields the smallest possible identity category with reliable data. The more fine-grained our measurement of the constituent parts of ethnic groups, the easier it is to bridge differences in group definitions between datasets.

The second step of linking ethnic categories is to exploit the structure of the linguistic tree. This tree is constructed by linguists based on the lexicographic similarity of any two languages/dialects and reflects the 'genealogy' of world languages (Desmet, Ortuño-Ortín, and Wacziarg 2012). The language tree helps us to assess the distances between any two languages, which proxy cultural (Fearon 2003) and genetic distances (Cavalli-Sforza 1997).

We illustrate the utility of linking different ethnic group lists via the language tree with an example from Ghana in Figure 2. Subfigure 2a depicts a simplified version of the subtree of the Akan language cluster in Ghana (black), which comprises the Abron and Akan languages as well as the Ahafo, Asante, and Fante dialects. To the right of the tree, we list four ethnic labels from four lists: the Akan from the Afrobarometer, the Asante/Akan from DHS, the Brong from Murdock's Map, and the Asante from the EPR data. In our approach, we link each of these labels to the relevant level on the language tree according to the similarity of the labels and other important clues such as demographic size and the presence or absence of any of its subsets/supersets. Any link to a higher-level language category implies a link to all subsidiary nodes below. Thus, linking the Akan

[3]Important exceptions are the Hutu and Tutsi in Rwanda and Burundi, Somali speaking clans in Somalia, as well as Arabic speaking groups in North Africa.

from the Afrobarometer to the 'Akan' node on level 9 simultaneously links them to all five language and dialect nodes below this node.

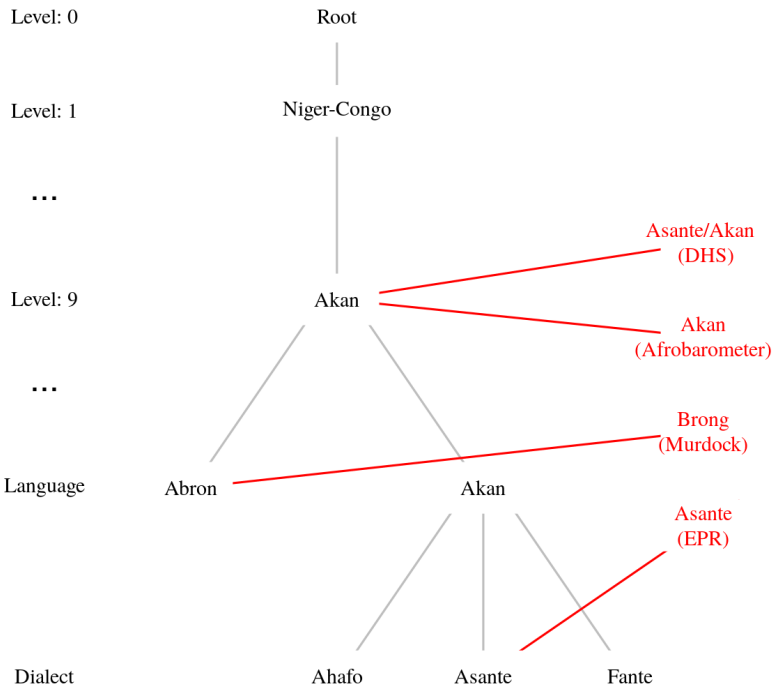
Once we establish the links between all datasets and the linguistic tree, we can merge any two datasets via three systematic rules. Researchers can adopt these rules according to their needs and fine-tune the trade-off between precision and completeness. When the goal is to achieve high levels of precision, researchers lose links between weakly related categories. Conversely, keeping as many groups as possible from one dataset comes at the cost of matching some groups in A to categories in B to which they are only weakly connected on the language tree.

Importantly, these links can be asymmetric, and thus allow us to connect multiple subgroups in A to a broader superordinate category in B without creating a reverse link. For example, researchers studying economic inequality between ethnic groups might measure groups' income from survey data and link it to an expert-coded list such as EPR. While the income estimates for large groups in EPR depends on correctly identifying all constituent survey groups, researchers might want to avoid income estimates for a small group in EPR from a large survey category, which would be biased from ecological inference.

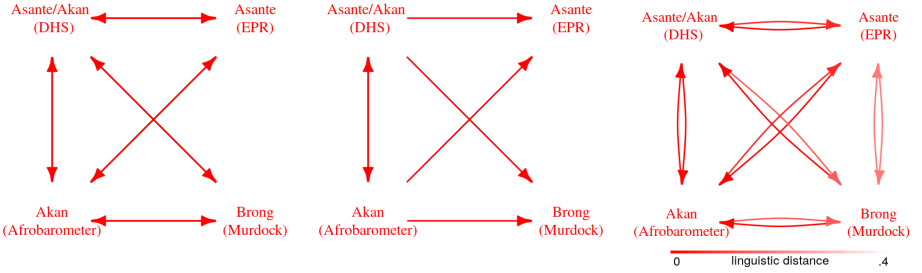
More concretely, we distinguish between three linking rules implemented in the LEDA R package:[4]

1. **Set overlap:** This rule generates a link between any two groups that share at least one language node at a specified level along the path that connects the tree's root and the node in question. In the example in Figure 2a, EPR's Asante and Murdock's Brong share the 'Akan' node at level 9 and all other nodes up to the root, whereas

[4]See the Appendix for details and example code.



(a) Language tree with links to ethnic categories from different datasets



(b) Set overlap: a and b share common dialect

(c) Share of common nodes: b covers 100% of a's dialects

(d) Linguistic distance

Figure 2. Partial linguistic tree from Ghana and link rules. Direction of arrows depicts direction of match: if $a \leftarrow b$, then b is matched to a but a is not matched to b .

Afrobarometer's Akan and EPR's Asante share all nodes from the 'Asante' node at the 'dialect' level to the root. We can now specify the trade-off between precision and completeness by choosing a level on which to match. Moving from the root to the dialect level increases precision but fails to connect some groups such as the Asante (EPR) and the Brong (Murdock) as shown in Figure 2b. We would match these two categories if linking groups via level 9 of the tree.

2. **Share of common nodes:** An alternative approach considers the degree of overlap between two ethnic categories at any given level of the language tree. Consider the level of dialects: EPR's Asante cover 1/4 of the dialects linked to Afrobarometer's Akan, while the latter cover all of the dialects linked to EPR's Asante. Once more, we face a trade-off between precision and completeness. Higher thresholds of nodes that two ethnic categories need to share generate fewer but more accurate links. For example, the most exact link for which group $b \in B$ must contain all dialects linked to a leads to asymmetric links. As Figure 2c demonstrates, only the Asante/Akan (DHS) and the Akan (Afrobarometer) have a reciprocal link since they correspond to the exact same nodes on the tree. While both are linked to the Brong (Murdock) and Asante (EPR), the latter are not linked back to them.

3. **Linguistic distance:** Finally, we can exploit the structure of the language tree to calculate the linguistic distance between any two groups linked to the tree. Doing so requires an assumption about how distances on the tree relate to observed linguistic distances. Following Fearon (2003), we can approximate the linguistic distance between two dialects or languages L_1 and L_2 as

$$1 - \left(\frac{2d(w(L_1, .., O) \cap w(L_2, .., O))}{d(w(L_1, .., O)) + d(w(L_2, .., O))} \right)^{\delta}$$

where $d(w(L_1, .., O))$ is the length of the path from the first language to the tree's origin and $d(w(L_1, .., O) \cap w(L_2, .., O))$ is the length of the intersection of the paths from the first and second language to the origin. Because we frequently match one ethnic category to several languages, we have to aggregate these distances, for example by taking the minimum distance between all dialects L_a in group a to any dialect L_b associated with group b. δ is an exponent to discount distances further away from the root of the tree. Figure 2d illustrates the resulting distances in our Ghanaian example. We can now define binary links by either specifying a linguistic distance threshold below which two groups are linked or linking each group to its closest linguistic neighbor(s). Alternatively, one may retain the continuous information of the distance measure and use e.g. the minimum linguistic distance between groups a and b for further analysis.

These three general rules allow for specifying the precision and coverage of links between any two group lists in a theoretically informed manner that reflects the needs of a research project. Researchers may also explore the impact of alternative linking rules by replicating their analyses across various ethnic links. Lastly, researchers can incorporate measures of uncertainty of any match into their analyses by, for example, weighting one-to-many matches between $a \in A$ and B by the linguistic distance between a and each category b to which it is linked.

Coding procedure and reliability

The quality of links between any two datasets depends on the quality of the links between any dataset and the Ethnologue list that we provide. We link 8'119 distinct ethnic categories from the eleven datasets in Figure 1 to the Ethnologue tree of African languages that

features 15'200 nodes, 2'154 primary languages and 4'822 dialects.[5] In this section, we describe our coding procedure, discuss the fit between different datasets, and show the results from different reliability tests.

To establish the link between a dataset and the Ethnologue tree, we follow a four-step procedure.[6] First, we use fuzzy string matching to create link suggestions between ethnic categories and Ethnologue entries. Second, we assign all ethnic group lists to human coders who suggest and justify links between ethnic categories and languages. The coders draw on the fuzzy string matches, information on groups' size, qualitative descriptions in codebooks, and secondary literature.

Third, we implement algorithms which ensure that links suggested by a coder actually exist in Ethnologue, and immediately transfer new links to ethnic categories with similar names in other datasets. While coders may still deviate from these automatic suggestions, e.g., when secondary sources suggest more plausible links, the procedure increases the consistency of our coding between different datasets. Fourth, we identify and check groups without a match, potentially inconsistent links between groups that share the same name, and inconsistent links between groups that cross borders.

To ensure the reliability of coded links, we implemented all four steps twice, and rotated coders between countries. Between these two rounds we recover 70% of all links precisely. Another 20% of all cases only differ by language tree level and still identify the same broader linguistic category. In about 4% of all cases, we link a language in one of the coding rounds but not in the other. In the remaining 5% of cases, we match ethnic categories to divergent sets of languages. This problem occurs most often in the AMAR dataset, which includes many highly disaggregated (historical) ethnic categories that are

[5]See also Appendix Table A1.
[6]We describe the details of our coding procedure in the online appendix.

hard to identify in Ethnologue. Finally, we double-checked the 30% of mismatches in a third round and decided on the optimal match.

After linking all ethnic datasets to Ethnologue, we can match ethnic categories from any two lists to each other. Appendix Figures A2 and A3 show that our language-based approach successfully links most ethnic categories from any specific dataset to at least one category in any other dataset. The share of successfully-linked groups decreases wherever we match fine-grained ethnic lists from census or survey data to more broadly defined groups (e.g. Murdock's map) or lists with theory-driven inclusion criteria and limited coverage (e.g. EPR and PREG).

Using this final list of links, we compared our coding to three links between the EPR dataset and the Afrobarometer, DHS, and Fearon's list (Cederman, Weidmann, and Bormann 2015), one link between EPR and DHS (Müller-Crepon and Hunziker 2018), and another between Murdock's map and the Afrobarometer (Nunn and Wantchekon 2011). Using the set overlap rule at the dialect level, we recover at least 90% of these earlier links between ethnic categories. Our recovery rate further increases as we link ethnic categories at lower levels on the tree.

Descriptive representation and trust in leaders

To illustrate the utility of our linking approach, we investigate whether African coethnicity with the president or other high-ranking government elites influences trust in state leaders. Analyses of whether, why, and under which conditions political loyalties follow ethnic cleavages figure prominently in the African politics literature (Bratton, Bhavnani, and Chen 2012; Ichino and Nathan 2013; Adida 2015). Positive assessments of co-ethnic government elites' in office may be rooted in cultural and psychological biases (Adida

et al. 2017), or reflect an instrumental logic of government favoritism towards co-ethnic constituencies (Bates 1974).

We combine data from Afrobarometer public opinion surveys on individual-level trust in the president with information from Vogt et al.'s (2015) EPR dataset on the representation of ethnic groups in government. We link individual Afrobarometer respondents via their spoken languages and the Ethnologue tree to the politically relevant ethnic groups in EPR. We construct binary links with the set overlap rule requiring that a respondent's language shares at least one node on level 16 (dialects) of the language tree with an EPR group (similar to Fig. 2b above). This choice reflects the fact that EPR contains a selected subset of ethnic categories with no claim to cover a country's entire population. Nearest linguistic neighbor or distance cutoff matches would risk linking EPR to ethnic categories that the creators of this dataset intentionally excluded from their list of politically relevant groups. With the set overlap matches, we construct dummy variables indicating, for each respondent, whether she is linked to an EPR group coded as at least government senior partner.[7] Additionally, we calculate respondents' linguistic distance to the closest EPR senior partner group or higher to measure their cultural proximity to government elites.

We then estimate linear models with country-survey and, in some specifications, ethnic group fixed effects along with common individual-level control variables (Table 1). In line with existing results, co-ethnicity with government officials increases trust in the president (Model 1). Results remain stable when only exploiting temporal changes in the ethnic composition of governments between survey rounds (Model 2), reducing the risk that our co-ethnicity variables capture unobserved differences between groups. Models 3 and 4

[7]EPR groups coded as senior partner or higher typically control the presidency or hold comparable shares of high-ranking government positions as the president's group.

TABLE 1 *Afrobarometer Analysis: Trust in President*

	Trust in President					
	Binary Link		Cont. Link		Both	
	(1)	(2)	(3)	(4)	(5)	(6)
Ethnic Link to Gov.	0.267***	0.259***			0.168***	0.198**
	(0.045)	(0.064)			(0.047)	(0.071)
Ling. Dist. to Gov			-0.363***	-0.374***	-0.218***	-0.163*
			(0.067)	(0.073)	(0.054)	(0.071)
Country-Survey FE	yes	yes	yes	yes	yes	yes
Ethnic Group FE	no	yes	no	yes	no	yes
Observations	141,674	141,674	137,543	137,543	137,543	137,543
Adjusted R^2	0.180	0.205	0.180	0.184	0.183	0.207

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

demonstrate that a smaller ethnic distance to ministers and government leaders similarly increases trust in presidents. Notably, we find distinct effects when introducing both variables into the same model. Both binary ethnic links and decreasing ethnic distance are associated with higher trust in African state leaders (Models 5 and 6) suggesting that cultural proximity to state elites matters beyond direct co-ethnicity.

Additionally, we conduct the same analysis with minister data from Francois, Rainer, and Trebbi (2015). Due to the temporal restrictions of their data, we retain just 6% of respondents from our original analysis. Nevertheless we still estimate statistically significant effects for Models 1 and 3, and substantive but statistically insignificant effects for Model 5.

Conclusion

In this research note, we introduce LEDA, a new tool to systematically link the most prominent datasets on African ethnic groups. The LEDA R package will facilitate research on the origins and consequences of ethnic identity in Africa and enable scholars to more rigorously analyze existing datasets. We hope that our contribution sparks new projects that exploit the potential of the linked data sources or add newly collected data. Our approach enables researchers to link their own ethnic group data to the language tree and then directly draw on information from all other linked datasets.

References

Adida, Claire L. 2015. "Do African Voters Favor Coethnics? Evidence From a Survey Experiment in Benin." *Journal of Experimental Political Science* 2 (1): 1–11.

Adida, Claire, Jessica Gottlieb, Eric Kramon, and Gwyneth McClendon. 2017. "Reducing or Reinforcing In-Group Preferences? An Experiment on Information and Ethnic Voting." *Quarterly Journal of Political Science* 12 (4): 437–477.

Afrobarometer. 2018. "Afrobarometer Data." *Available at http://www.afrobarometer.org*.

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

Birnir, Johanna K, Jonathan Wilkenfeld, James D Fearon, David D Laitin, Ted Robert Gurr, Dawn Brancati, Stephen M Saideman, Amy Pate, and Agatha S Hultquist. 2014. "Socially Relevant Ethnic Groups, Ethnic Structure, and AMAR." *Journal of Peace Research* 52 (1): 110–115.

Bratton, Michael, Ravi Bhavnani, and Tse-Hsin Chen. 2012. "Voting Intentions in Africa: Ethnic, Economic or Partisan?" *Commonwealth & Comparative Politics* 50 (1): 27–52.

Cavalli-Sforza, L Luca. 1997. "Genes, peoples, and languages." *Proceedings of the National Academy of Sciences* 94 (15): 7719–7724.

Cederman, Lars-Erik, Nils Weidmann, and Nils-Christian Bormann. 2015. "Triangulating Horizontal Inequality: Toward Improved Conflict Analysis." *Journal of Peace Research* 52 (6): 806–821.

De Luca, Giacomo, Roland Hodler, Paul A Raschky, and Michele Valsecchi. 2018. "Ethnic Favoritism: An Axiom of Politics?" *Journal of Development Economics* 132:115–129.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg. 2012. "The Political Economy of Linguistic Cleavages." *Journal of Development Economics* 97 (2): 322–338.

DHS. 2018. "Demographic and Health Surveys." *Integrated Demographic and Health Series (IDHS), version 2.0, Minnesota Population Center and ICF International. Available at http://idhsdata.org.*

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

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

Francois, Patrick, Ilia Rainer, and Francesco Trebbi. 2015. "How is Power Shared in Africa?" *Econometrica* 83 (2): 465–503.

Gellner, Ernest. 1983. *Nations and Nationalism.* Ithaca, NY: Cornell University Press.

Huber, John D. 2012. "Measuring Ethnic Voting: Do Proportional Electoral Laws Politicize Ethnicity?" *American Journal of Political Science* 56 (4): 986–1001.

Ichino, Nahomi, and Noah L. Nathan. 2013. "Crossing the Line: Local Ethnic Geography and Voting in Ghana." *American Political Science Review* 107 (2): 344–61.

Lewis, M. Paul, ed. 2009. *Ethnologue: Languages of the world.* Vol. 16. SIL International Dallas, TX.

McCauley, John F. 2014. "The Political Mobilization of Ethnic and Religious Identities in Africa." *American Political Science Review* 108 (4): 801–816.

Minnesota Population Center. 2017. *Integrated Public Use Microdata Series, International: Version 6.5.* Minneapolis, MN: University of Minnesota. doi:http://doi.org/10.18128/D020.V6.5.

Müller-Crepon, Carl, and Philipp Hunziker. 2018. "New Spatial Data on Ethnicity: Introducing SIDE." *Journal of Peace Research* 55 (5): 687–698.

Murdock, George Peter. 1959. *Africa. Its Peoples and Their Culture History.* New York: McGraw-Hill Book Company.

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

Østby, Gudrun. 2008. "Polarization, horizontal inequalities and violent civil conflict." *Journal of Peace Research* 45 (2): 143–162.

Posner, Daniel N. 2004a. "Measuring Ethnic Fractionalization in Africa." *American Journal of Political Science* 48 (4): 849–863.

———. 2004b. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98 (4): 529–545.

Robinson, Amanda Lea. 2014. "National versus Ethnic Identification in Africa: Modernization, Colonial Legacy, and the Origins of Territorial Nationalism." *World Politics* 66 (4): 709–746.

Vail, LeRoy. 1989. *The Creation of Tribalism in Southern Africa.* Berkeley, CA: University of California Press.

Vogt, Manuel, Nils-Christian Bormann, Seraina Ruegger, Lars-Erik Cederman, Philipp Hunziker, and Luc Girardin. 2015. "Integrating Data on Ethnicity, Geography, and Conflict: The Ethnic Power Relations Dataset Family." *Journal of Conflict Resolution* 59 (7): 1327–1342.

Weidmann, Nils B., Jan Ketil Rød, and Lars-Erik Cederman. 2010. "Representing Ethnic Groups in Space: A new dataset." *Journal of Peace Research* 47 (4): 491–499. doi:10.1177/0022343310368352. http://dx.doi.org/10.1177/0022343310368352.

Wig, Tore. 2016. "Peace From the Past: Pre-Colonial Political Institutions and Civil Wars in Africa." *Journal of Peace Research* 53 (4): 509–524.

Online Appendix to

Linking Ethnicity in Africa: Data and Methods

TABLE OF CONTENTS

DESCRIPTIVES

TABLE A1 *Matched ethnic group lists*

List	Countries	Groups	Groups by country	Geo data	Source
Afrobarometer	36	1582	43.9	Point	Survey
AMAR	50	1560	31.2	—	Expert
DHS	29	1471	50.7	Point	Survey
Ethnic Power Relations [EPR]	53	298	5.6	Polygon (0/1)	Expert
Ethnologue	53	2409	45.5	Polygon (0/1)	Expert
Fearon	48	361	7.5	—	Expert
Francois, Trebbi & Rainer [FRT]	15	279	18.6	—	Expert
GREG	52	491	9.4	Polygon (0/1)	Expert
IPUMS	15	639	42.6	Raster (%)	Census
Murdock Map	50	1310	26.2	Polygon (0/1)	Expert
PREG	41	128	3.1	—	Expert
SIDE	23	499	21.7	Raster (%)	Survey

Note: Because of spelling inconsistencies, groups in the Afrobarometer, DHS, IPUMS, and SIDE lists include 'duplicate' entries. Groups that span multiple countries are counted multiple times.

Coding Procedure

The language-based link between any two ethnic group datasets requires that each ethnic category in these two lists (Table A1) are mapped to the language(s) and language families associated with the group. We link about 8'100 distinct ethnic categories[8] 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 A2 *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 A2 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.[9] All of these automatic matches are generated via fuzzy

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

[9]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

string matching,[10] and represent suggestions of decreasing quality. As Table A2 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.

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 A2), or descriptions of the groups in the respective codebooks.[11] 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 A4 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 A5 lists all unique

country the coder is working on.

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

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

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 A2. 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.[12] 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' 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 somewhat later in the process and only coded once.

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

RELIABILITY

Table A3 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 A3 *Intercoder reliability: By list type*

Type	N	Equal	Partial overlap	Missing match	Disjoint
All	8,068	0.70	0.20	0.04	0.05
Afrobarometer	1,582	0.78	0.13	0.05	0.05
AMAR	1,606	0.71	0.17	0.04	0.08
DHS/SIDE	1,489	0.76	0.14	0.06	0.04
EPR	311	0.59	0.30	0.09	0.03
Fearon	361	0.70	0.22	0.04	0.04
FRT	279	0.67	0.29	0.01	0.03
GREG	491	0.70	0.26	0.004	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 the Afrobarometer and DHS surveys, one link between EPR and Fearon's list (Cederman, Weidmann, and Bormann 2015), one link between EPR and DHS (Müller-Crepon and Hunziker 2018), and a final link between Murdock's Map and the Afrobarometer (Nunn and Wantchekon 2011).[13] 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).[14]

We recover matches in existing link files in at least 90% of all cases at the highest resolution, i.e., the dialect level.[15] 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.

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

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

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

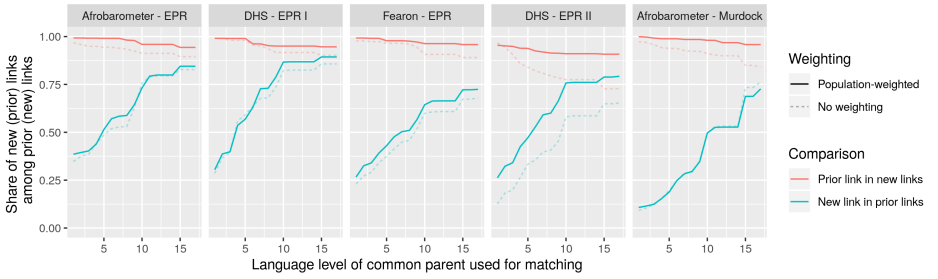


Figure A1. Recovery of previously coded links between groups by matching groups via common parent nodes at varying Ethnologue language levels (see Figure 2)

Fit between datasets

Here, we provide a brief overview of (1) the links between our lists of group categories and Ethnologue's language tree and (2) the resulting links between any of the lists of ethnic categories.[16]

The first column in Figure A2 plots the population-weighted shares of groups matched to Ethnologue with darker squares indicating better matches.[17] We match almost 100% of the population-weighted ethnic categories in all lists to the language tree. The lowest match rate is achieved for the DHS and SIDE data with 97.8%.

After linking each list to the language tree, we can now link all lists to each other via shared Ethnologue nodes. For the following comparisons, we match ethnic categories in a

[16]For these analyses, we drop all ethnic categories which we have identified as not relating to ethnic groups or identities. The categories only occur in individual-level survey data and refer, for example, to towns and other locations, regions, or even, at times, surnames.

[17]Matching a large group such as the Asante in Ghana has more weight than matching a small group with only 1'000 members. All matches are weighted relative to all other groups in the respective country and list.

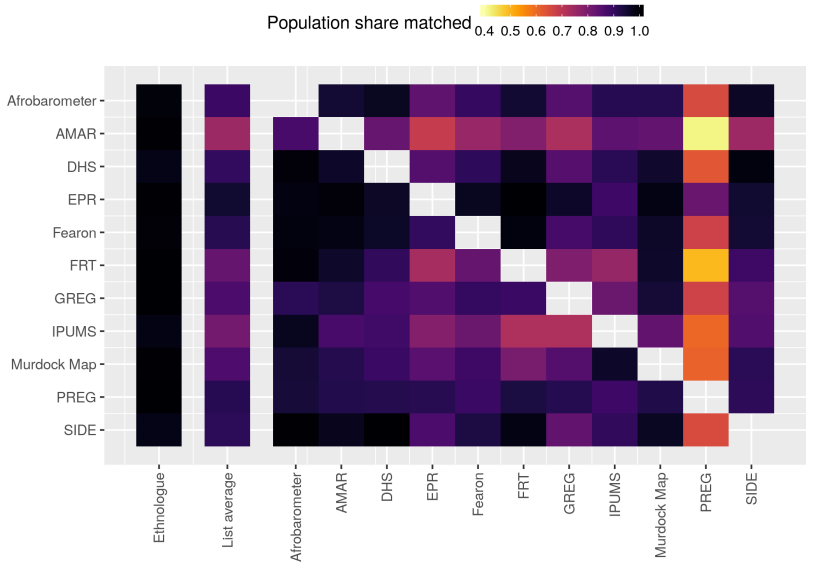


Figure A2. Proportion of groups per list matched to Ethnologue and other lists. Groups are weighted according to their size proportional to the population of their country. Population weights are calculated by country(-year, e.g. in surveys). Figures either come from the data sets directly or are calculated by using the GRUMP population data (Linard et al. 2012) for the year 2000 aggregated at the ethnic polygon level. AMAR and PREG do not come with population figures or data on the spatial extent of ethnic groups, which is why groups from these lists receive the same weights.

non-restrictive way that resembles the matches employed in previous studies (e.g. Cederman, Weidmann, and Bormann 2015; Nunn and Wantchekon 2011). More specifically, we match any two groups as soon as they are linked to at least one common dialect (cf. Figure 2b). We do not impose any minimum threshold requiring a certain proportion of shared dialects or maximum linguistic distance between linked groups.

For each ethnic list pair A and B, we calculate the share of ethnic categories $a \in A$ that are linked to at least one category in B, again weighting by a by country-specific population shares as described above. The second column in Figure A2 displays the average of this

calculation for each individual list *A* (individual rows) across all other lists *b* (columns 3 to 13). For example, almost all EPR ethnic groups have matches in most other lists whereas many ethnic categories from AMAR and IPUMS have no links to groups from other lists. The size of ethnic categories offers one explanation for these differences. The EPR dataset contains fairly large, politically relevant ethnic groups. These broad ethnic categories are likely to have at least one counterpart in any other dataset. Conversely, datasets with comprehensive lists of fine-grained ethnic categories such as AMAR and IPUMS will feature lots of groups with no linguistic link to the selected set of groups listed, e.g. in EPR and PREG. Yet this is unlikely to explain the weaker performance of the Murdock, GREG/Atlas Narodov Mira, and Francois, Rainer & Trebbi's lists. One important reason for the relatively low overall matching share of between 87 and 89 percent may be their age (for the first two) and their different conceptualization of ethnic categories.

The remaining columns (3-13) in Figure A2 encode the population share in *A* (row) successfully matched to list *B* (column). This perspective reveals how the choice of baseline ethnic categories matters for the ability to make connections between two datasets. Consider the Afrobarometer to EPR link (row 1, column 5) and the EPR to Afrobarometer link (row 4, column 3). We only match around 85% of the fine-grained ethnic categories enlisted in the Afrobarometer survey data to EPR groups. In contrast, we can match essentially all EPR categories to at least one group from the Afrobarometer list. Match rates between the different lists are on average lower once we look at groups without weighting them according to their population (see Figure A3 in the Appendix). The lower recovery rate results from fewer matches between datasets with lots of small groups (AMAR, IPUMS, DHS) and datasets with larger ethnic categories such as the EPR and GREG.

Our baseline matching succeeds in linking groups *a* in *A* and *b* in *B* via a large

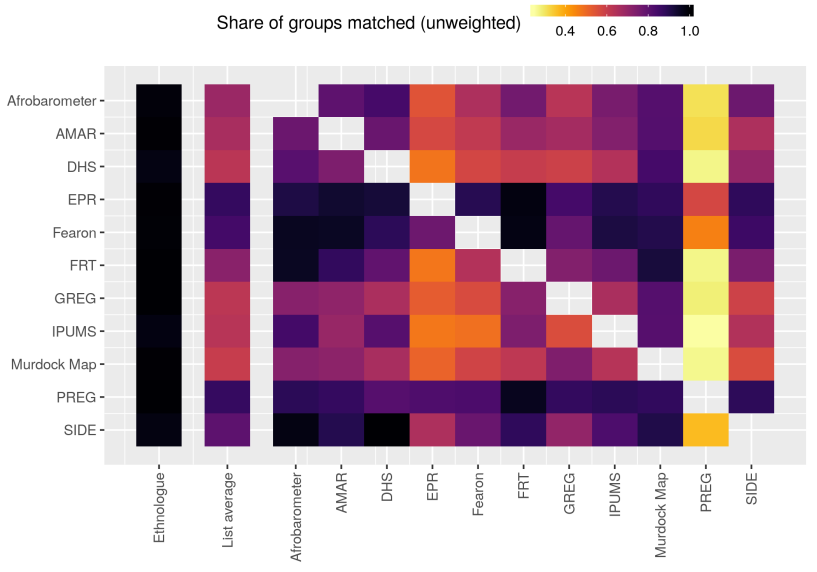


Figure A3. Proportion of groups per list matched to Ethnologue and other lists. Each ethnic category receives the same weight.

proportion of dialects associated with groups *a*. This is equivalent to the matching criterion depicted in Figure 2b discussed above. On average, groups *a* are linked via 93% of the dialects associated with them to groups *b*. Conversely, the dialects of group *a* include, on average, 83% of the dialects associated with groups *b*. These summary statistics show that our baseline matching often produces results in which group *a* is a subset of the groups *b* it is matched to. This of course mostly occurs when we match very fine-grained ethnic categories from from survey or census dataset *A* to aggregate dataset *B* such as EPR or GREG.

Additional Tables

TABLE A4 *Overlap between coded and automatically proposed matches*

Type	Matches coded			Matches proposed		
	Total	same as proposed match (in %)		Total	same as coded match (in %)	
		Org. name	Any		Org. name	Any
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 automatically proposed match. Thus, in the case of Afrobarometer, of 1560 matches, 25% have been proposed automatically based on the name of an Ethnologue language. 50% have been proposed based on any name, alternative name or subdialect of a language, or language from other countries. Reversely, of the 3724 proposals made for Afrobarometer matches, only 21% have been coded as actual match.

TABLE A5 *Groups without a match in Ethnologue*

Country	Groups
AGO	Kakondas [AMAR]; KOROCA [Murdock Map]
BFA	Kibsi [AMAR]
BWA	Mokgothu [Afrobarometer]; Sekgothu [Afrobarometer]
CAF	Besom [AMAR]
CIV	Eda [AMAR]; French [Afrobarometer]
CMR	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]
COG	Bahumbu [DHS]; Bakaya [DHS]; Bweni [DHS]; IKASA [Murdock Map]; Kabinda [DHS]; Mayanga [DHS]; Minkengue [DHS]
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]
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]; Mophiring [Afrobarometer]; Motaung [Afrobarometer]; Motsoeneng [Afrobarometer]; Mzema [Afrobarometer]; Sephotsa [Afrobarometer]
MDG	Tealaotra [Afrobarometer]; Zaza lava mahafasa [Afrobarometer]
MLI	Trouka [Afrobarometer]
MOZ	Islamic Coastal [Fearon]
MUS	Muslims [EPR]

NGA	Agazawa [DHS]; Ahu [DHS]; Amamong [DHS]; Awo [DHS]; Bafeke [DHS]; Bagathiya [DHS]; Bageri [DHS]; Bagunge/Badagire [DHS]; Bahnake [DHS]; Barabaci [DHS]; Beteer [DHS]; Buko [DHS]; Chiba [DHS]; Dumak [DHS]; Eterco [Afrobarometer]; Gmenchi [DHS]; Gomo/Gamoyaya [DHS]; Gumbarawa [DHS]; Igbanko [DHS]; Ijeme [DHS]; Ikara [DHS]; Knale [Afrobarometer]; Kuba [Afrobarometer]; Kunkawa/Kawa [DHS]; Mbwa [DHS]; Mgas [Afrobarometer]; Mirnang [DHS]; Muryan [DHS]; Nanba/Wanba [Afrobarometer]; Nezou [DHS]; Nkwana [Afrobarometer]; Nnebe [DHS]; Normana [Afrobarometer]; Obubua [DHS]; Ohari [DHS]; Omele [DHS]; Paibun [DHS]; Pasama [DHS]; Rulere [DHS]; Sekere [DHS]; Somunka [DHS]; Taira [DHS]; Tangoa [Afrobarometer]; Uhionigbe [DHS]; Uru [DHS]; Yendre [DHS]; Yonubi [DHS]
TCD	Kanem-Bornou [DHS]
TGO	Ndebele [Afrobarometer]
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	Shangaan/Tsonga/Ronga/Tswa [Afrobarometer]
ZMB	American [DHS]; American [IPUMS]; European [DHS]; European [IPUMS]; North-Western [DHS]
ZWE	Vhitori [Afrobarometer]

TABLE A6 *Trust in President: EPR & FRT*

	Trust 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

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", "NIG"),
                                        round = 4, marker = "language"),
                                 lists.b = list(type = c("FRT"),
                                                iso3c = c("UGA", "NIG")),
                                 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: K
##   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 authors /at/ xxxx 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, and Nils-Christian Bormann. 2015. "Triangulating Horizontal Inequality: Toward Improved Conflict Analysis." *Journal of Peace Research* 52 (6): 806–821.

Linard, Catherine, Marius Gilbert, Robert W Snow, Abdisalan M Noor, and Andrew J Tatem. 2012. "Population distribution, settlement patterns and accessibility across Africa in 2010." *PloS one* 7 (2): e31743.

Müller-Crepon, Carl, and Philipp Hunziker. 2018. "New Spatial Data on Ethnicity: Introducing SIDE." *Journal of Peace Research* 55 (5): 687–698.

Nunn, Nathan, and 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.