Linking Ethnic Data from Africa

Carl Müller-Crepon,[*] Yannick Pengl, and Nils-Christian Bormann

Abstract

Social scientists in general and conflict researchers in particular increasingly combine multiple datasets to study ethnic politics and conflict 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. An analysis of political exclusion, mistrust of state leaders, and ethnic grievances highlights different ways of linking ethnic categories from multiple datasets. The LEDA open-source software package allows researchers to link ethnic groups from any database with explicit rules and to add their own data on ethnic groups.

Keywords: Ethnicity, Africa, Data

Word count: 5,998

[*] Corresponding author: carl.mueller-crepon@icr.gess.ethz.ch

Introduction

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 & Wantchekon, 2011), voting (Huber, 2012), and distributive politics (De Luca et al., 2018). Ethnic groups and their attributes have been especially relevant to the study of civil war (Horowitz, 1985; Stewart, 2008; Østby, 2008; Cederman, Gleditsch & Buhaug, 2013) and communal violence (Fjelde & von Uexkull, 2012; Fjelde & Østby, 2014; Hillesund et al., 2018), but also one-sided violence (Fjelde & Hultman, 2014) and international dynamics of ethnic civil wars (Cederman et al., 2013). Combining meso- and micro-level datasets, scholars explore the effects of ethnic group-level characteristics on individual outcomes (Franck & Rainer, 2012), measure group-level attributes through micro-data (Cederman, Weidmann & Bormann, 2015), or enrich one meso-level dataset with information from another (Wig, 2016; Wig & Kromrey, 2018).

When studying questions related to ethnicity, it is inherently difficult to link ethnic categories from two datasets to each other.[1] Due to the socially constructed nature of ethnic identities and different conceptual approaches, we lack a common definition of the universe of ethnic groups in Africa. 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),

[1] The terms 'linking' and 'matching' interchangeably denote the process of connecting any two ethnic categories from different data sources.

focus on a theoretically motivated subset of 'politically relevant' ethnic categories, including group clusters with multiple ethnic identities. Others, such as the All Minorities at Risk data (AMAR; Birnir et al., 2014), identify as many 'socially relevant' categories as possible. Individual-level data such as the Demographic and Health Surveys (DHS, 2018) identify respondents' language. As a result, ethnic categories from different datasets do not easily map onto one another.

In this article, we introduce the Linking Ethnic Data from Africa (LEDA) project. We match more than 8,100 ethnic categories from the eleven most prominent datasets on ethnic groups in Africa to the list of known language families, languages, and dialects from the 16th edition of the Ethnologue database (Lewis, 2009). Using the Ethnologue linguistic tree as a relational master dictionary allows 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, within and across country borders. Figure 1 depicts our approach with the datasets linked to each other. Appendix Table A3 provides additional information on the inclusion criteria and substantive contents of these datasets.

LEDA aims to improve empirical research on ethnicity by increasing efficiency, transparency, and conceptual clarity of linking ethnic data. First, scholars who merge two datasets hard-code several decisions into their data, such as the resolution at which groups are linked or the required degree of overlap between two groups. LEDA allows scholars to explore the robustness of their results to these decisions. Second, matching tables are often not accessible to other researchers,

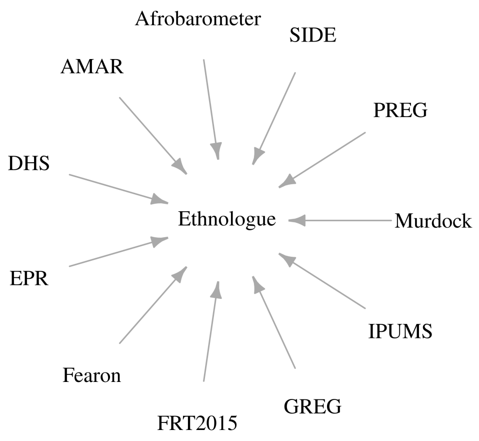


Figure 1: Meta-structure of the dictionary approach.
Data sources: Afrobarometer (2018); AMAR: Birnir et al. (2014); DHS (2018); EPR: Cederman, Wimmer & Min (2010); Fearon (2003); FRT: Francois, Rainer & Trebbi (2015); GREG: Weidmann, Rød & Cederman (2010), IPUMS: Minnesota Population Center (2017);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

which limits replication attempts. Third, the current fragmentation of links between group lists makes it difficult to leverage the information they contain for linking new group lists to existing ones. With LEDA, researchers who want to establish new ethnic links can draw on the information contained in all prior links.

The open-source LEDA R package[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. This flexibility permits scholars to draw on the large pool of ethnic group-level data when working with geographic or survey data. Thus, LEDA increases the number and scope of research questions that can be studied with currently available and newly collected data on ethnic groups in Africa.

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

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 $b \in B$ contained in two different datasets. First, the two datasets might 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, their ethnic categories are not necessarily 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 ideally goes beyond a binary link logic 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. Most social science definitions stress subjective beliefs in common descent or (descent-based) membership criteria as defining features of ethnic as opposed to other social groups (Weber, 1978; Barth, 1969; Chandra, 2012). Although individuals in Africa subscribe to multiple putatively descent-based identities including tribe, religion, and race (Posner, 2004b; McCauley, 2014), language is arguably the most wide-spread ethnic identity marker globally (Gellner, 1983), and is particularly pronounced in Sub-Saharan Africa due to, not least, missionary activity (Vail, 1989). More importantly, other ethnic markers often closely align with language. In many African states, language mirrors tribal affiliations at the local level, yielding the smallest 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. Our purely language-based approach is insufficient in contexts where non-linguistic categories are more salient than or further divide linguistic ones.[3] Future research may extend LEDA by adding ethnic categories such as religion or race to the matching dictionary.

The second step of linking ethnic categories leverages 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 (e.g., Gray & Atkinson, 2003). The language tree helps us to assess the distances between any two languages, which proxy cultural (Fearon, 2003) and genetic distances (Cavalli-Sforza, 1997).

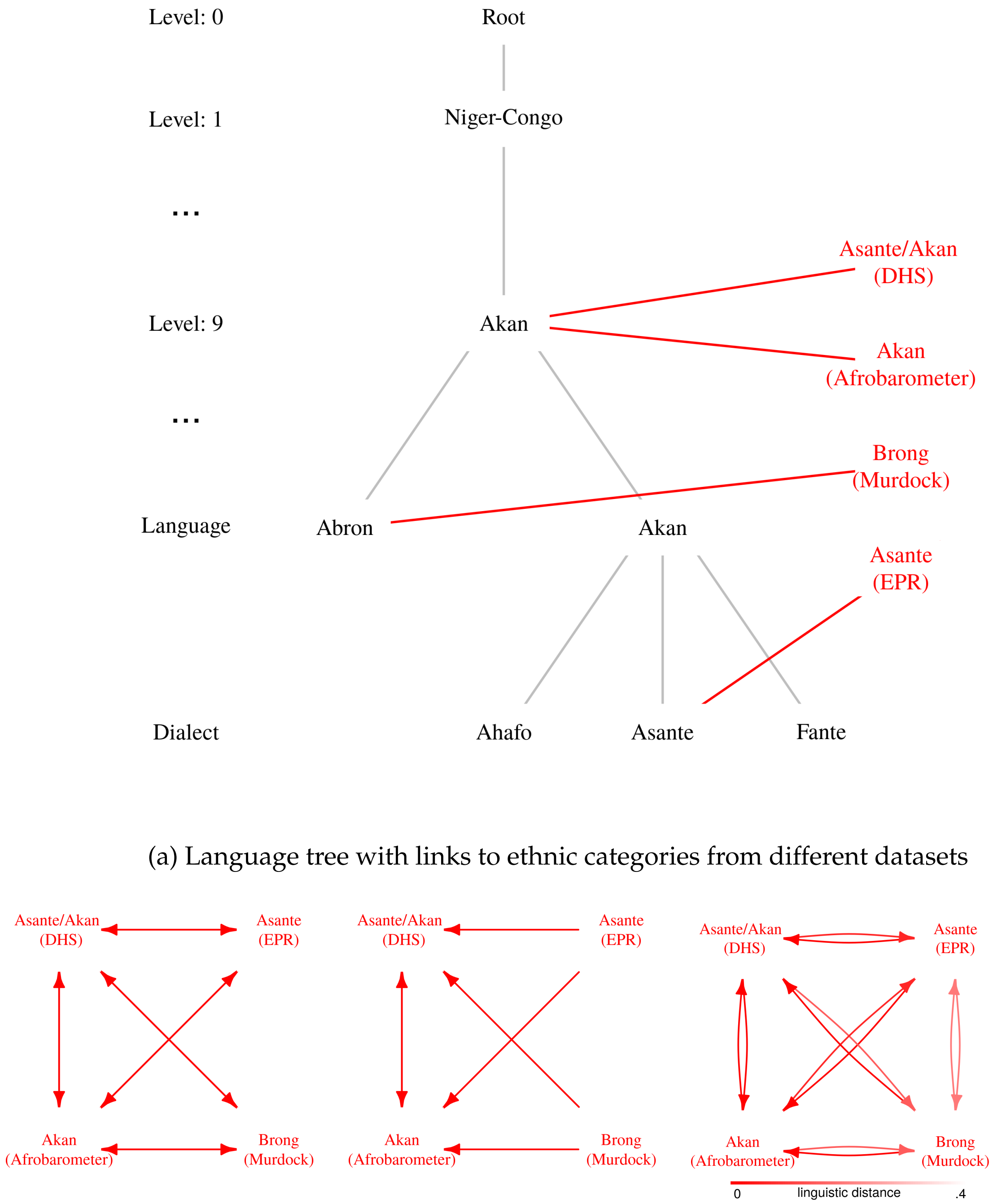
We illustrate the utility of linking different ethnic group lists via the language tree with an example from Ghana in Figure 2. Subfigure 2a depicts the simplified subtree of the Akan language cluster in Ghana (black), comprising the Abron and Akan languages as well as the Ahafo, Asante, and Fante dialects. 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. 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

[3] Important African cases are the Hutu and Tutsi in Rwanda and Burundi and Somali speaking clans in Somalia.

as demographic size and information from datasets' codebooks. In cases in which the appropriate tree-level link was ambiguous, we gave preference to more encompassing links, i.e. linking the Akan to the Akan language cluster rather than to the Akan language. Any link to a higher-level language category implies a link to its subsidiary nodes. Thus, linking the Akan from the Afrobarometer to the 'Akan' node on level 9 simultaneously links them to the language and dialect nodes below.

Once we have linked all datasets to 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 will encounter some groups for which no precise links exist. Conversely, keeping as many groups as possible from one dataset comes at the cost of matching groups that are only weakly related.

Importantly, these links can be asymmetric, connecting multiple subgroups in B to a broader superordinate category in A 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 depend 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 comprises many respondents from other ethnic categories than the narrow one that EPR identified as politically relevant.



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

(b) ”Set overlap: link ethnic labels a from list A and b from list B if a and b share a common dialect.”

(c) Share of common nodes: link ethnic label b from list B to label a in target list A if and only if a covers 100% of b's dialects.

(d) Linguistic distance

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

More concretely, we distinguish between three linking rules. These are implemented in the LEDA R package, which is documented in the Appendix:

1. **Set overlap:** This rule generates a link between any two groups that share at least one language node at a specified level of the language tree. 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 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; Figure 2b). A link via level 9 of the tree would match these two categories.

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 a must contain all dialects linked to b 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. Both the Brong (Murdock) and

Asante (EPR) are linked to the superordinate Akan (Afrobarometer) and Asante/Akan (DHS), but no reverse links exist.

3. **Linguistic distance:** Finally, we can use the language tree to calculate the linguistic distance between any two groups. Following Fearon (2003), we can approximate the linguistic distance between two dialects or languages L_1 and L_2 as the fraction of their paths to the tree root that they share.[4] 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 languages L_a in group a to any dialect L_b associated with group b. 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. Alternatively, the continuous information of the distance measure, e.g., the minimum linguistic distance between groups a and b, can serve for further analysis.

These three general rules allow for specifying the precision and coverage of links between any two group lists within or across countries 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 un-

[4] Mathematically, linguistic distances are thus calculated as:
$D_{L_1,L_2} = 1 - \left(\frac{2d(w(L_1,..,O) \cap w(L_2,..,O))}{d(w(L_1,..,O)) + d(w(L_2,..,O))}\right)^{\delta}$,
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. δ is an exponent to discount distances further away from the root of the tree; it is typically set to .5.

certainty of any match into their analyses by weighting one-to-many matches by the linguistic distance between group a and linked categories b.

Coding procedure and reliability

The quality of links between any two datasets depends on the quality of their links to the Ethnologue dateset. The main challenge is to correctly match different names or spellings that describe the same category. We link 8,119 distinct ethnic categories from the eleven datasets in Figure 1 and Table I to the Ethnologue tree of African languages that features 15,200 nodes, 2,154 primary languages and 4,822 dialects.

Table I: Matched ethnic group lists

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

To establish the link between a dataset and the Ethnologue tree, we follow a four-step procedure.[5] First, we use fuzzy string matching to create link suggestions between ethnic categories and Ethnologue entries and their alternative names. Second, we assign all ethnic group lists to research assistants who code and justify links between ethnic categories and language tree nodes. The coders draw on the fuzzy string matches, information on groups' size, qualitative descriptions in codebooks, and secondary sources containing ethnonyms, spoken languages, and other relevant information.

Third, an algorithm checks that coded links actually exist in Ethnologue and adds new links as suggestions for ethnic categories with similar names in other datasets. This procedure increases the consistency of our coding across different datasets, while allowing coders to deviate from these automatic suggestions, e.g., when secondary sources suggest more plausible links. Fourth, we check groups without a match, potentially inconsistent links of groups that share the same name, and inconsistent links of groups that cross borders.

To ensure reliability of our coding decisions, we repeated these four steps, and rotated coders between countries. Between the two rounds we recover 70% of all links. Signaling difficulties in determining the 'resolution' of ethnic groups, 20% of all cases differ by language tree level but identify the same broader linguistic category for a group. 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 hard to identify in Ethnologue. Finally, the authors double-

[5] We describe additional details of our coding procedure in Online Appendix A .

checked the 30% of mismatches in a third round and decided on the optimal match based on the comments and sources provided by our coders and, where necessary, additional investigation.

Moreover, we compare the links between ethnic group lists derived from our coding to three links between the EPR dataset and the Afrobarometer, DHS, and Fearon's list (Cederman, Weidmann & Bormann, 2015), one link between EPR and DHS (Müller-Crepon & Hunziker, 2018), and another between Murdock's map and the Afrobarometer (Nunn & 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 results of ethnic group links

After linking all ethnic datasets to Ethnologue, we can match ethnic categories from any two lists to each other. Figure 3 shows 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.

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, weighting categories a by their population shares.[6] The first column of Figure 3 shows that the match-rate be-

[6] Note that we drop all obviously non-ethnic group labels ('others', 'don't know', etc.) for the following analysis. For equivalent non-weighted results, see Appendix Figure A2.

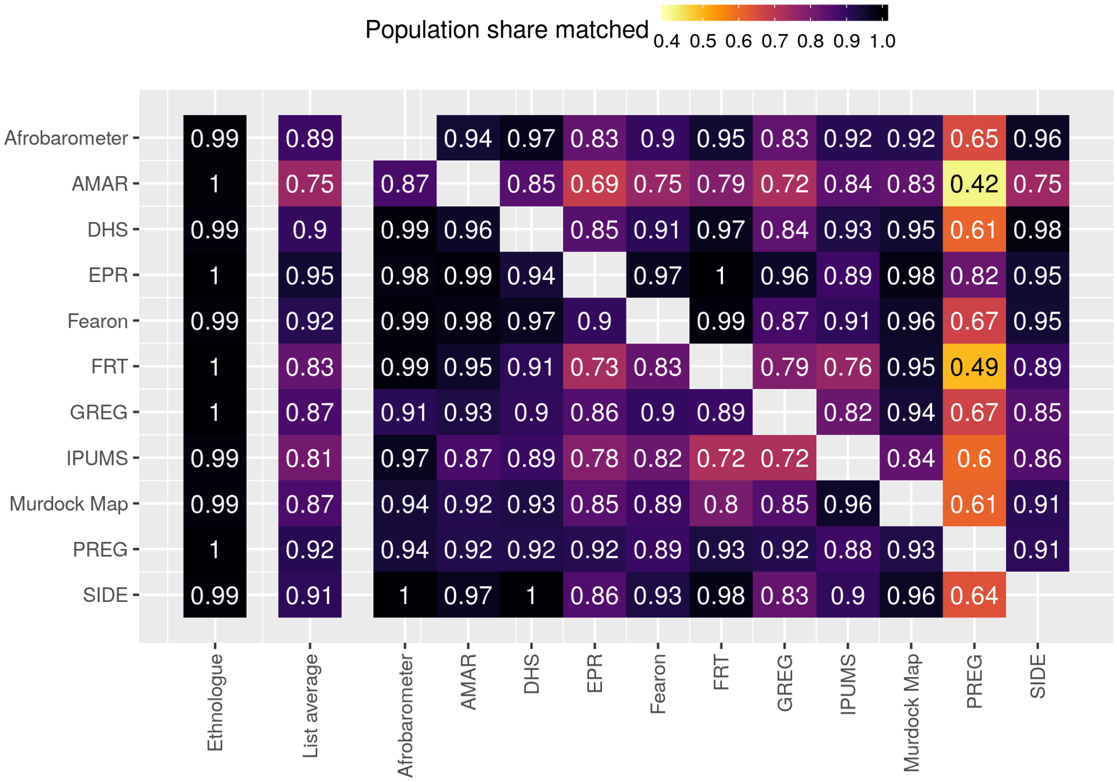


Figure 3: Proportion of groups matched to Ethnologue and other lists. Groups are weighted according to their size proportional to the population of their country.
Note: Population weights are calculated by country(-year, e.g. in surveys). Figures come from the original data or are geographically retrieved from the 2000 GRUMP data (Linard et al., 2012). Data from AMAR and PREG is processed without population weighting.

tween the average ethnic category lists A (rows) and the Ethnologue data is 99% or above. The second column displays the average match rate of lists A across all other lists b (columns 3 to 13).[7] For example, 95% of the population of EPR groups is recovered in the average list B whereas only 81% of the population in the IPUMS census data is matched to groups in B, on average. The granularity 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,

[7] For results by country see Table A4.

many fine-grained ethnic categories have no link to the selected sets of groups listed, e.g., in EPR and PREG. Other important reasons for variation in matched population shares are list age (Murdock and GREG) and different conceptualizations of ethnic categories (e.g. FRT vs. AMAR).

The remaining columns (3-13) in Figure 3 encode the population share of groups a (row) successfully matched to groups b (columns). This disaggregation 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 6) and the EPR to Afrobarometer link (row 4, column 3). We only match around 83% of the fine-grained ethnic categories enlisted in the Afrobarometer survey data to EPR groups. In contrast, we match essentially all EPR categories to at least one group from the Afrobarometer. Without population weighting match rates decrease because of fewer matches between many small groups in fine-grained datasets (AMAR, IPUMS, DHS) and groups in datasets with large ethnic categories (EPR and GREG) (see Figure A2 in the Appendix).

Different types of errors arise due to missing links between ethnic group lists and the language tree. Two broad classes of false negatives exist. First, some definitions of ethnic categories do not have linguistic equivalents in Ethnologue. For example, we could not find a suitable match for the religiously defined 'Muslims' in EPR's group list of Mauritius.[8] Second, some non-matches occur because the list of languages is too detailed. It is often difficult to identify all the constituent languages of big ethnic clusters. For example, we probably miss some of the links between the EPR cluster 'Hausa-Fulani and the Muslim Middle Belt' in Nigeria

[8] Refer to Table A6 in the appendix for a list of all non-matches by country and dataset.

and the hundreds of corresponding Ethnologue languages, many of which have a few thousand speakers only. Conversely, false positives also exist. They affect links between ethnic groups wherever two groups speak the same language but differ along other historical, phenotypical, or religious markers. Important examples include the Hutu and Tutsi in Burundi and Rwanda, as well as Arab and Somali-speaking groups. Researchers should take note of such cases and correct language-based links accordingly.

Empirical illustration

To illustrate the utility of LEDA, we investigate whether exclusion from political power leads African citizens to distrust their political leaders and develop ethnic grievances. While the empirical link between ethnic exclusion and intrastate conflict is well established at the ethnic group-level (see e.g. Cederman, Wimmer & Min, 2010), only few, inconclusive findings on the micro-foundations of the underlying processes exist. Most importantly, it remains contested whether individuals reflect objective ethno-political inequalities in perceived injustice and grievances (Hillesund et al., 2018). We use our ethnic links to test whether group-level political exclusion affects subjectively felt distrust of those in power and perceptions of ethnic discrimination as is often assumed in the conflict literature.

We combine information from Vogt et al.'s (2015) EPR dataset on the representation of ethnic groups in government with data from Afrobarometer Afrobarometer (2018) surveys on respondents' mistrust in state leaders and their perceptions of ethnic discrimination by the government.[9] After linking respondents via their language and the Ethnologue tree to the politically relevant ethnic groups in EPR, we construct binary measures of political representation as well as continuous linguistic distances to the most powerful ethnic group(s). Citizens may react more strongly to 'foreign rule' by an ethnically distant elite than a more proximate one.

First, we use the set overlap rule requiring that a respondent's language shares at least one node on the dialect level of the language tree with an EPR group (see Fig. 2b above). We then construct dummy variables indicating, for each respondent, whether she is linked to an EPR group coded as at least government senior partner.[10] Second, we calculate respondents' linguistic distance to the closest EPR senior partner group or higher to measure their cultural proximity to the most high-ranking 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 (Tables II and III). In line with existing theories, co-ethnicity with government senior partner increases trust in the president (Model 1 in Table II). The estimates imply .25 points greater mistrust on a standardized scale between 0 and 1 among less represented groups. Results remain stable when only exploiting temporal changes in the ethnic composition of governments between survey rounds

[9] The respective survey items are: (1) 'How much do you trust each of the following, or haven't you heard enough about them to say: The President/Prime Minister?' and (2) 'How often is [Respondent's Ethnic Group] treated unfairly by the government?' We standardize responses to a mean of 0 and standard deviation of 1.

[10] EPR groups coded as senior partner or higher control the presidency or hold comparable shares of high-ranking government positions as the president's group. Individuals receive a 0 if they either belong to an EPR group coded as junior partner, politically powerless or discriminated against, or 'politically irrelevant.' The latter category comprises all ethnic groups in a country that are not part of the EPR dataset.

Table II: Afrobarometer analysis: Mistrust in president

	Mistrust 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

(Model 2), reducing the risk that our co-ethnicity variables capture unobserved differences between groups. Models 3 and 4 demonstrate that larger linguistic distances to the most powerful ethnic groups similarly increase mistrust in leaders. Notably, we find separate effects when introducing both variables into the same model (Models 5 and 6) suggesting that cultural distance to political power matters beyond direct co-ethnicity.

Table III: Ethnic grievances: Unfairly treated by government

	Unfair treatment of own group					
	Binary Link		Cont. Link		Both	
	(1)	(2)	(3)	(4)	(5)	(6)
Ethnic Link to Gov.	-0.337^{***}	-0.089			-0.200^{***}	-0.024
	(0.047)	(0.052)			(0.059)	(0.076)
Ling. Dist. to Gov			0.474^{***}	0.463^{***}	0.303^{***}	0.175
			(0.080)	(0.089)	(0.079)	(0.143)
Country-Survey FE	yes	yes	yes	yes	yes	yes
Ethnic Group FE	no	yes	no	yes	no	yes
Observations	123,650	123,650	119,546	119,546	119,546	119,546
Adjusted R^2	0.143	0.174	0.138	0.145	0.142	0.170

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

Results for the more direct measure of ethnic grievances about unfair treatment by the government are substantively similar but somewhat weaker (Table III). The linguistic distance results appear more robust to the inclusion of group fixed effects than our binary measure of political representation (Models 1–4) and the estimates in Model 6 lose statistical significance. Additionally, we conduct the same analysis with data on leaders' ethnicity from Francois, Rainer & 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 and very similar effects if we include only the binary or continuous ethnic representation measure. Estimates from models including both terms show the same pattern but fail to reach significance. (Tables A7 and A8).

Overall, these results are consistent with the notions that (1) exclusion from power translates into distrust and grievances among ordinary citizens and (2) that ethnic dominance by culturally distant elites may spur even stronger frustration than exclusion from power per se. Our findings thus provide novel evidence for the first step of the causal chain that links ethnic inequality in political representation to conflict via widespread grievances among members of disadvantaged groups.

Conclusion

In this article, we introduce LEDA, a new tool that systematically links 11 datasets on African ethnic groups to each other. The LEDA R package facilitates research on the origins and consequences of ethnic identity in Africa and enables scholars to make the most out of existing datasets. Our approach and technical infrastructure also enables researchers to link their own ethnic group data – for example on the ethnic identities of violent actors and their victims – to the language tree and directly combine it with information from all other linked datasets.

More generally, the LEDA project presents a versatile solution to the grouping problem of ethnic identities that permeates existing datasets. As different lists of ethnic groups are based on differing definitions of ethnic identities, linking them becomes cumbersome and oftentimes involves non-replicable, arbitrary decisions. Drawing on the tree of languages as a 'dictionary,' LEDA helps researchers who combine various datasets to address the grouping problem of ethnic identities in a transparent and replicable manner. While currently based on linguistic markers among ethnic groups in Africa, the approach is generally extendable to other world regions and ethnic markers.

Replication data

The R-package and code for the empirical analysis in this article, along with the Online appendix, can be found at http://www.prio.org/jpr/datasets and https://github.com/carl-mc/LEDA. All analyses have been conducted using R 3.4.

Acknowledgements

We thank Cullen Hendrix and two anonymous referees for excellent feedback and Paola Galano Toro, Vanessa Kellerhals, Benjamin Füglister, Lukas Dick, Carlos Mairoce, and Julian Seitlinger for invaluable research assistance. We are furthermore grateful for comments from Levke Aduda, Matthew Gichohi, and participants of the 2018 AFK Workshop in Hamburg and the 2019 APSA Annual Meeting.

Funding

This research was supported by the Swiss National Science Foundation through the grants P0EZP1_165233 (Carl Müller-Crepon) and P0EZP1_159076 (Yannick Pengl).

References

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

Barth, Fredrik (1969) Ethnic groups and boundaries - introduction. *In Fredrik Barth (Ed.) Ethnic Groups and Boundaries: The Organization of Cultural Differences. Little, Brown: Boston*: 9–37.

Birnir, Johanna K; Jonathan Wilkenfeld, James D Fearon, David D Laitin, Ted Robert Gurr, Dawn Brancati, Stephen M Saideman, Amy Pate & Agatha S Hultquist (2014) Socially relevant ethnic groups, ethnic structure, and AMAR. *Journal of Peace Research* 52(1): 110–115.

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

Cederman, Lars-Erik; Kristian S. Gleditsch & Halvard Buhaug (2013) *Inequality, Grievances, and Civil War*. New York, NY: Cambridge University Press.

Cederman, Lars-Erik; Kristian Skrede Gleditsch, Idean Salehyan & Julian Wucherpfennig (2013) Transborder ethnic kin and civil war. *International Organization* 67(2): 389.

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

Cederman, Lars-Erik; Andreas Wimmer & Brian Min (2010) Why do ethnic groups rebel? new data and analysis. *World Politics* 62(1): 87–119.

Chandra, Kanchan (2012) *Constructivist Theories of Ethnic Politics*. Oxford, UK: Oxford University Press.

De Luca, Giacomo; Roland Hodler, Paul A Raschky & Michele Valsecchi (2018) Ethnic favoritism: An axiom of politics? *Journal of Development Economics 132*: 115–129.

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.

Fjelde, Hanne & Lisa Hultman (2014) Weakening the enemy: A disaggregated study of violence against civilians in Africa. *Journal of Conflict Resolution* 58(7): 1230–1257.

Fjelde, Hanne & Gudrun Østby (2014) Socioeconomic inequality and communal conflict: A disaggregated analysis of Sub-Saharan Africa, 1990–2008. *International Interactions* 40(5): 737–762.

Fjelde, Hanne & Nina von Uexkull (2012) Climate triggers: Rainfall anomalies, vulnerability and communal conflict in Sub-Saharan Africa. *Political Geography* 31(7): 444–453.

Franck, Raphael & 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 & 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.

Gray, Russell D & Quentin D Atkinson (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965): 435–439.

Hillesund, Solveig; Karim Bahgat, Gray Barrett, Kendra Dupuy, Scott Gates, Håvard Mokleiv Nygård, Siri Aas Rustad, Håvard Strand, Henrik Urdal & Gudrun Østby (2018) Horizontal inequality and armed conflict: A comprehensive literature review. *Canadian Journal of Development Studies/Revue canadienne d'études du développement* 39(4): 463–480.

Horowitz, Donald L. (1985) *Ethnic Groups in Conflict*. University of California Press.

Huber, John D (2012) Measuring ethnic voting: Do proportional electoral laws politicize ethnicity? *American Journal of Political Science* 56(4): 986–1001.

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

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

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.

Müller-Crepon, Carl & 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 & 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.

Posner, Daniel N. (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.

Stewart, Frances (2008) *Horizontal Inequalities and Conflict: Understanding Group Violence in Multiethnic Societies*. London, UK: Palgrave Macmillan.

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 & Luc Girardin (2015) Integrating data on ethnicity, geography, and conflict: The ethnic power relations dataset family. *Journal of Conflict Resolution* 59(7): 1327–1342.

Weber, Max (1978) *Economy and Society*. Los Angeles, CA: University of California Press.

Weidmann, Nils B; Jan Ketil Rød & Lars-Erik Cederman (2010) Representing ethnic groups in space: A new dataset. *Journal of Peace Research* 47(4): 491–499.

Wig, Tore (2016) Peace from the past: Pre-colonial political institutions and civil wars in Africa. *Journal of Peace Research* 53(4): 509–524.

Wig, Tore & Daniela Kromrey (2018) Which groups fight? Customary institutions and communal conflicts in Africa. *Journal of Peace Research* 55(4): 415–429.

Authors

CARL MÜLLER-CREPON, b. 1989, PhD in Political Science (ETH Zurich, 2019); Postdoctoral Researcher, International Conflict Research, ETH Zurich, currently visiting Harvard University.

YANNICK PENGL, b. 1987, PhD in Political Science (ETH Zurich, 2018); Postdoctoral Researcher, International Conflict Research, ETH Zurich.

NILS-CHRISTIAN BORMANN, b. 1985, PhD in Political Science (ETH Zurich, 2014); Professor of International Political Studies, Witten/Herdecke University. [Senior Lecturer, University of Essex before 10/2020]

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[11] 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.[12] All of these automatic matches are generated via fuzzy string matching,[13] 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.

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

[12] 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.

[13] 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.[14] 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.[15] 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 de-

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

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

cisions 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 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 ex-

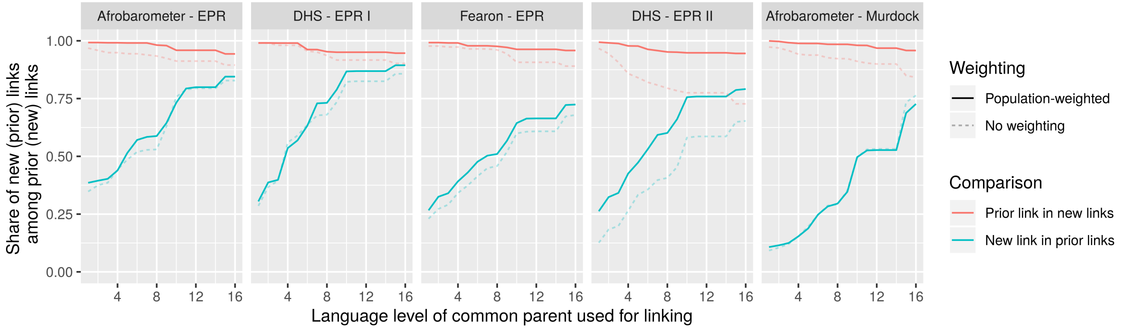


Figure A1: Recovery of previously coded links between groups by matching groups via common parent nodes at varying Ethnologue language levels (see Figure 2)

isting 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 & 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).[16] 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).[17]

We recover matches in existing link files in at least 90% of all cases at the highest resolution, i.e., the dialect level.[18] 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.

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

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

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

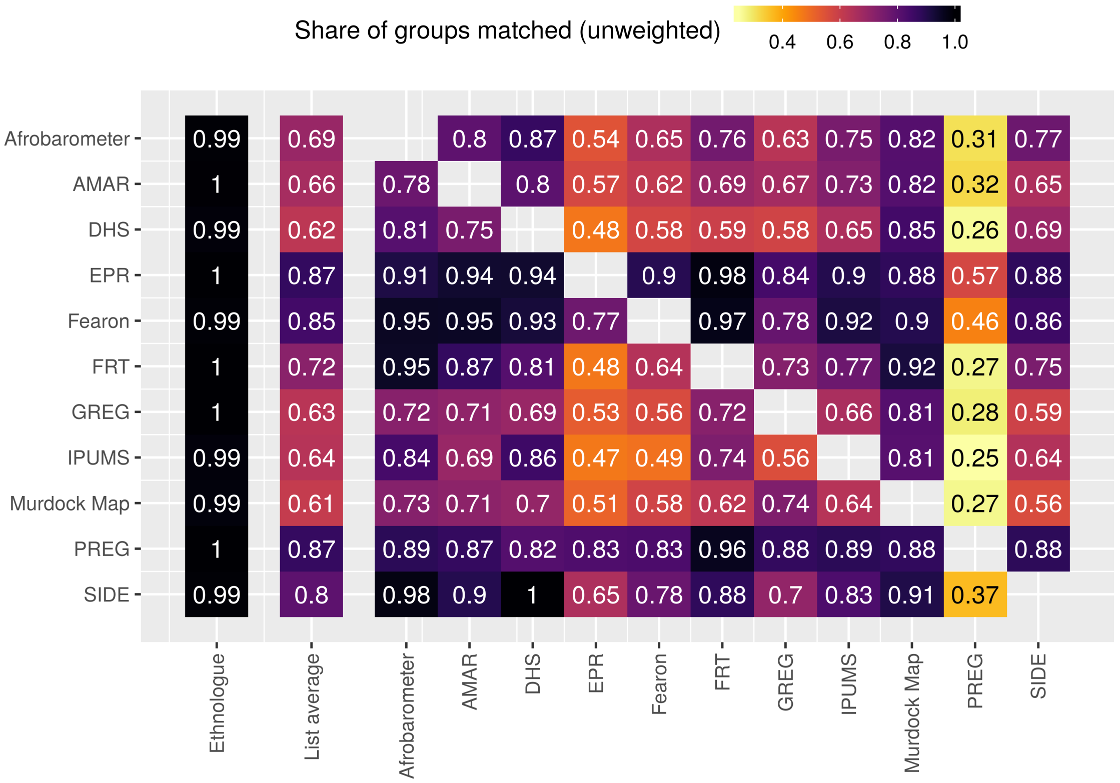
C Additional Figures and Tables

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	



	Ethnologue	List average	Afrobarometer	AMAR	DHS	EPR	Fearon	FRT	GREG	IPUMS	Murdock Map	PREG	SIDE
Afrobarometer	0.99	0.69		0.8	0.87	0.54	0.65	0.76	0.63	0.75	0.82	0.31	0.77
AMAR	1	0.66	0.78		0.8	0.57	0.62	0.69	0.67	0.73	0.82	0.32	0.65
DHS	0.99	0.62	0.81	0.75		0.48	0.58	0.59	0.58	0.65	0.85	0.26	0.69
EPR	1	0.87	0.91	0.94	0.94		0.9	0.98	0.84	0.9	0.88	0.57	0.88
Fearon	0.99	0.85	0.95	0.95	0.93	0.77		0.97	0.78	0.92	0.9	0.46	0.86
FRT	1	0.72	0.95	0.87	0.81	0.48	0.64		0.73	0.77	0.92	0.27	0.75
GREG	1	0.63	0.72	0.71	0.69	0.53	0.56	0.72		0.66	0.81	0.28	0.59
IPUMS	0.99	0.64	0.84	0.69	0.86	0.47	0.49	0.74	0.56		0.81	0.25	0.64
Murdock Map	0.99	0.61	0.73	0.71	0.7	0.51	0.58	0.62	0.74	0.64		0.27	0.56
PREG	1	0.87	0.89	0.87	0.82	0.83	0.83	0.96	0.88	0.89	0.88		0.88
SIDE	0.99	0.8	0.98	0.9	1	0.65	0.78	0.88	0.7	0.83	0.91	0.37	

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`.

```r
## 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.

```r
## 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.

```r
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.

```r
## 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.

E References

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.