African Studies

ISSN: 0002-0184 (Print) 1469-2872 (Online) Journal homepage: https://www.tandfonline.com/loi/cast20

Linguistic Situation in Twenty sub-Saharan African Countries: A Survey-based Approach

Katalin Buzasi

To cite this article: Katalin Buzasi (2016) Linguistic Situation in Twenty sub-Saharan African Countries: A Survey-based Approach, African Studies, 75:3, 358-380, DOI: 10.1080/00020184.2016.1193376

To link to this article: https://doi.org/10.1080/00020184.2016.1193376

AFRICAN STUDIES, 2016
VOL. 75, NO. 3, 358–380
http://dx.doi.org/10.1080/00020184.2016.1193376

Routledge
Taylor & Francis Group

Linguistic Situation in Twenty sub-Saharan African Countries: A Survey-based Approach

Katalin Buzasi

Utrecht University

ABSTRACT

Data on second languages in sub-Saharan Africa are hard to come by. Consequently, any source that contributes to our knowledge beyond the level of primary languages should be appreciated and exploited. This article utilises Round 4 of the Afrobarometer Survey that collects information on ethnicity, home, and additional languages in 20 sub-Saharan African countries. The study has three main contributions. First, it overviews and compares some widely used sources that contain linguistic data and investigates why they show such a diverse picture on language use patterns. Second, it applies the ICP which, according to the author's knowledge, is the first linguistic measure that takes multilingualism into account. Third, it shows how a simple graphic representation of the ICP can be used to visualise the most important dimensions of a country's linguistic situation including the order of languages according to their size, the presence of monolingual speakers, and the relation between vernaculars and the former colonisers' languages. The study findings are expected to be of interest to scholars engaged in language policy and planning and language-related development issues.

ARTICLE HISTORY
Received 29 January 2015
Accepted 21 April 2015

KEYWORDS
sub-Saharan Africa; linguistic situation; multilingualism; linguistic diversity; linguistic data; Index of Communication Potential; Afrobarometer Survey

It is well established that Africa is characterised by high ethnic and linguistic heterogeneity (Lewis, Simons & Fennig 2014; Alesina et al. 2003; Academija nauk SSSR 1964), multilingual citizens (Lewis et al. 2014; Laitin 2007), and high risk of language death especially in areas close to the Equator (Nettle & Romaine 2000). However, one finds oneself in a difficult situation when it comes to actual numbers to describe the aforementioned dimensions of the linguistic situation. While population censuses and certain surveys (for instance the Demographic and Health Surveys, DHS) usually provide information on ethnicity, mother tongue or home language, obtaining data on additional languages (an essential requirement for analysing the patterns of multilingualism and language dynamics) is more difficult.

This article attempts to fill this lack within the literature to a certain extent by utilising Round 4 of the Afrobarometer Survey (2008 & 2009)[1] that contains not only the ethnicity and home language but also the additional languages of more than 27,000 individuals in 20 sub-Saharan African countries. Although the review of the size of ethnic and linguistic

CONTACT Katalin Buzasi K.Buzasi@uu.nl
Supplemental data for this article can be accessed http://dx.doi.org/10.1080/00020184.2016.1193376

groups and the distribution of other than home languages would already be a substantial contribution to our understanding of the language patterns in sub-Saharan Africa, this article aims to present findings in a more insightful way. This study predominantly relies on the Index of Communication Potential (ICP) (Buzasi 2015) that has several advantages as a linguistic indicator. First, according to the author's knowledge, it is the first linguistic measure that accounts for multilingualism and is calculated for multiple countries. Second, since it builds on individual language repertoires, the ICP can be applied to visualise the most important dimensions of a country's linguistic situation including the order of languages by their size, the presence of monolingual speakers within linguistic groups, the relationship between European and indigenous languages, and the number of languages in the typical citizen's repertoire. Moreover, this study overviews some of the most widely used linguistic data sources and investigates why they provide a diverse picture on the linguistic situation in sub-Saharan Africa. The results are expected to be of interest to language and education planners; economic and political scientists focusing on the development consequences of diversity and multilingualism; and other researchers whose work includes designing and interpreting surveys including questions on language or ethnicity.

The following section of this article gives an overview of the use of linguistic data; the next section discusses the benefits and limitations of the Afrobarometer Survey as a linguistic data source and its comparison with other available materials; the ICP is then introduced and is followed by a graphical representation of the ICP to show the patterns of language use in the 20 sample countries in a comparative way; finally, the article concludes and discusses how the findings relate to other disciplines.

The Use of Linguistic Data

Linguistic information is collected for a number of purposes. While the World Atlas of Language Structures (WALS) is concerned with the structural (phonological, grammatical, and lexical) properties of languages, the Open Language Archives Community (OLAC) aims to collect available material in and on the languages of the world.[2] However, since they are more relevant from the aspects of this study, I discuss the sources that cover Africa (as a continent) or African countries and contain quantitative data on the size and use of languages.

Large databases that attempt to represent and understand the patterns of multilingualism across the world define the first central area where linguistic data are essential. The Ethnologue (Lewis et al. 2014), one of the main materials of this article, and the *Atlas of the World's Languages* (Asher & Moseley 2007) serve as general reference catalogues. Both classify, list, and map languages by country, provide information on the number of speakers, and compile language-specific bibliographies. In order to support Christian missionary activities and to measure the share of 'unreached' people, the Joshua Project also collects data on the size of ethnic and linguistic groups.[3] Since they are proper sources to estimate linguistic diversity (Desmet, Ortuno-Ortin & Wacziarg Forthcoming; Fearon 2003), the aforementioned three databases are extensively used in development studies, economics, and political science. The empirical literature has established that economic growth (Easterly & Levine 1997; Pool 1972), social capital (Putnam 2007; Alesina & La Ferrara 2002), and the quality of government (Mauro 1995) are negatively associated with, while the

probability of internal conflicts (Montalvo & Reynal-Querol 2005, 2010) is positively associated with (ethno)linguistic diversity.

The second area which requires information on the number of speakers and their geographical concentration is the field of language policy and planning. The 'Survey of language use and language teaching in Eastern Africa' conducted in 1968–1971 financed by the Ford Foundation and sponsored by local universities was the first large-scale sociolinguistic research initiative in Africa.[4] The project resulted in five volumes containing the classification and size, the historical and socioeconomic context (including attitudes), and the educational role of languages in Ethiopia, Kenya, Tanzania, Uganda, and Zambia (references are presented in the online supplementary material). The *Language and Dialect Atlas of Kenya* edited by Bernd Heine and Wilhelm Möhlig in the 1980s had a similar objective. Systemic language surveying projects have recently been implemented in South Africa and Tanzania. The five language atlases (references are provided in the online supplementary material and in Van der Merwe & Van der Merwe 2006: 1–2), which utilise South African census data from 1980, 1991, and 2001, provide information on the national and regional distribution of the official languages and the socioeconomic characteristics (for example religion, age structure, education, segregation index) of their speakers. An additional language project was initiated by UNESCO in the late 1990s (UNESCO 2000). The primary goal of the 'Languages of Tanzania' project launched in 2001 at the University of Dar es Salaam was to promote local languages which are not officially recognised. The outcomes of the project include several lexicons, dictionaries, and a language atlas (Chuo Kikuu cha Dar es Salaam 2009) that presents the number of L1, L2, L3 speakers of local languages at the national, provincial and district levels (a detailed overview of the challenges and results of the project is provided in Muzale & Rugemalira 2008). Comprehensive articles describing the sociolinguistic situation and evaluating the language policy of Botswana, Malawi, Mozambique, South Africa, Algeria, Côte d'Ivoire, Nigeria, and Tunisia are published in various issues of *Language Policy and Planning in Africa* (Baldauf & Kaplan 2004) and *Language Planning and Policy in Africa* (Kaplan & Baldauf 2007). The role of indigenous languages in education is one of the most debated issues (see for example Rabenoro 2013; Capo, Gbeto & Huannou 2009). Some works address specific language policy questions such as the violation of language rights (Namyalo & Nakayiza 2014).

And finally, collecting data on language use behaviour is especially important in the case of minority and endangered languages. The speakers of small languages, which usually lack official recognition in Africa, are more prone to poverty (Harbert et al. 2009). Certain languages have more social, cultural, economic, and political value than others (Batibo 2005: 93–4). Theories explaining language death agree that if the expected benefits from identifying with another language are high enough, people are likely to abandon their language of origin (Mesthrie et al. 2009: 248–51: Fishman 1991). Hence, the loss of speakers is recognised as a sign of increased endangerment (Lewis & Simons 2010; UNESCO 2003; Fishman 1991). Having acknowledged the social problems associated with language death and linguistic diversity loss, numerous programmes have been initiated to identify threatened languages (namely *UNESCO Atlas of the World Languages in Danger* (Moseley 2010) and the Catalogue of Endangered Languages, under the direction of University of Hawai^{c}i at Mānoa and LINGUIST List/Eastern Michigan University) and to reverse the process of language decline across the world (for example language

development works at SIL International, projects within the UNESCO Endangered Languages Programme and projects financed by the US National Science Foundation).[5]

The Afrobarometer Survey as a Linguistic Data Source

The survey

The Afrobarometer Survey is an independent, non-political research initiative to map the social, political, and economic atmosphere in Africa. Since it is conducted regularly[6] and provides a representative sample of citizens of voting age[7] and a standard set of questions, the Afrobarometer has recently become an acknowledged source of development-related research (Eifert, Miguel & Posner 2010; Nunn 2010). Unfortunately, additional languages are included only in Round 4; thus, this article is limited to the 20 countries[8] and the two consecutive years (2008 and 2009) covered in that wave. In this article, the ethnic and linguistic situation is measured by three AB variables: Q3 (Which Ghanaian/Kenyan/etc language is your home language?); Q79 (What is your tribe? You know, your ethnic or cultural group); and Q88E (What languages do you speak well?). While respondents were required to select their ethnicity and home language from a predefined list, languages in Q88E are completely based on self-report. The 16th edition of *Ethnologue* (Lewis 2009) is employed to identify languages when they are referred to by alternate names.

Benefits

Beyond providing information on the complete language repertoire, using the Afrobarometer for describing the linguistic situation has several other advantages.

First, the basic units of the survey are individuals. Unlike sources that report only the share of the population speaking certain languages as primary or secondary (for instance Lewis et al. 2014), individual level data allow a country's typical language repertoire to be captured, to identify linguistic groups that tend to remain monolingual, and to spot which languages are complementaries or substitutes. In addition, individual level data can be aggregated at any desired level of analysis (country, region, urban-rural distinction etc), thus can still be applied for studies with a macro-level approach.

Second, the Afrobarometer covers 20 out of the 54 African countries in a single conceptual framework. Since the sampling method and the surveying period is the same for all countries, African societies can be compared based on a comprehensive source where observed differences across countries cannot be assigned to the diversity in the applied methodologies. For instance, the *Ethnologue* (Lewis et al. 2014) often reports the number of language speakers within a country based on sources from different years or even decades. The case of Namibia serves as an illustration: while most data are taken from a source from 2006 (which is not specified), the number of Naro and !Xóõ speakers is based on Maho (1998) and Traill (1985), respectively. Moreover, data on second languages provided by *Ethnologue* (Lewis et al. 2014) are quite incidental and their sources are not always reported correctly. The only way to obtain data on other than home languages is to browse available country and sociolinguistic reports and to handle the differences in the data collecting methods.

Third, ethnicity and languages are surveyed separately in Afrobarometer. Although it is logical to assume that these two concepts are identical or at least greatly similar, Africa provides several cases where this is not the case. Development studies often proxy ethnicity with linguistic data when information on the former is not available (Cheeseman & Ford 2007). The Afrobarometer helps to reveal how large the gap between the size of an ethnic, and the corresponding linguistic, group can be and why.

Limitations

The Afrobarometer, however, has two obvious limitations as a linguistic source. First, the codebook and the questionnaire do not define 'tribe', 'ethnicity', and 'well-spoken languages' and in the case of 'home language', the manual is rather confusing. Defining and measuring the aforementioned terms are among the main concerns of several disciplines including sociolinguistics, second language acquisition, anthropology, and political science. Second, minority and endangered linguistic groups are underrepresented in the Afrobarometer. According to the sampling manual, the survey occasionally purposely oversamples certain populations that are politically significant within a country to ensure that the size of the sub-sample is large enough to be analysed.

Q3 explicitly intends to collect information on home languages. In case the respondent does not understand the question completely, the questionnaire suggests the following 'clarification sentence': 'That is, the language of your group of origin'. This is a confusing choice of words though. While home language is usually understood as the language most frequently spoken at home, the clarification sentence seems to refer to a different linguistic concept, namely first language, which is usually defined as the language that a person learns first in childhood (Gass & Selinker 2008: 7, Chuo Kikuu cha Dar es Salaam 2009: xii). Although the two concepts are often considered as synonyms in everyday use, due to migration, inter-ethnic marriage and language shift, the language spoken at home with spouses, children and relatives might be different from one's first language. While population censuses and other surveys generally collect information on home languages (see the online supplementary material), linguistic research (bilingual and multilingual studies, language teaching and second language acquisition) rather works with the 'first language-second language-(third language)-etc' distinction.

Ethnicity, surveyed in Q79, is a hotly debated multidimensional concept which is difficult to measure (Brown & Langer 2010; Burton, Nandi & Platt 2010; Hale 2004). While some view ethnic identity as a stable sense of group belonging based on common biological origin, historical experiences, traditions, culture, and language (Horowitz 1985), others argue that ethnic identity is a fluid concept which is often used as a tool by the elites to mobilise the population in economic and political competition (Banton 1997). As a result, it can be manipulated by certain means even in the short-run: empirical research has shown that the proximity of political elections intensifies ethnic group identification in Africa (Eifert et al. 2010).

Q88E ('Which languages do you speak well?') has two main shortcomings: it does not indicate what 'speaking well' means and is completely based on self-report. Although it is the field of language teaching and second language acquisition where measuring language proficiency is the most relevant, population censuses and other demographic surveys also contain information on certain linguistic abilities occasionally (the next

section and the online supplementary material discuss this issue in more detail). However, linguistic and non-linguistic surveys differ greatly in terms of depth, the covered areas of competencies and the applied evaluation methods. Linguistic surveys differentiate and cover various fields of abilities such as reading, writing, listening, and speaking, and the knowledge of grammar and vocabulary (Alderson 2005); and carefully design the test and the scoring system in order to gain a refined picture on the learners' achievements (North 2000). In contrast, language-related questions in demographic surveys are regularly less specified and not adequately elaborated from linguistic aspects. Censuses and demographic surveys usually focus on literacy, a key aspect of human capital and human development, and rely on self-assessment or very simple evaluation techniques (for example reading a simple sentence on a card). Without measuring it or offering several proficiency categories at least, it is difficult to tell the actual level of proficiency in languages listed in Q88E.

Political scientists often argue:

> People are very bad reporters of their own language repertoires – some lie (especially to political authorities) about their competency in certain languages; others are simply unaware of the languages (or speech forms) they use in different contexts. (Laitin 2000: 144)

The prestige of languages and the respondents' sociocultural identity might also encourage one to report a language one does not speak adequately or to suppress the ones one commands (Laitin 2000; Baetens Breadsmore 1982). Linguists highlight that reported and measured language proficiency differ for a number of reasons other than political. Anxiety and experience with the language under question (the number of years spent with learning the language, failure in linguistic test) can bias self-assessment (MacIntyre, Noels & Clément 1997). Moreover, it is also possible that the test is not adequately designed and does not mirror real abilities (Pray 2005).

The aforementioned limitations make it necessary to specify the linguistic terms used in this study. Languages in Q3 are referred to as home languages, and groups in Q79 as ethnic groups. Languages listed in Q88E are referred to as additional languages or other than home languages, but are not labelled as second languages. There are various reasons for doing so. First, linguistics generally applies the concept of second language as the complementary of first language: second language can be any language learned after the first language (Ortega 2009: 5–7). Thus, using second language along with the concept of home language would be a divergence from the usual practice. Second, since Q88E contains all the languages that the respondent speaks without any further clarification, calling them simply second language would raise additional issues, for example the distinction between second and foreign languages (Gass & Selinker 2008: 7) or the distinction between second, third and additional languages (Ortega 2009: 5–7), which will not be addressed here. And third, since our study aims to focus on the multilingual nature of sub-Saharan African societies in the first place without any intention to contribute to the debate on the aforementioned linguistic terms, the more flexible label of 'additional languages' or 'other than home language' is enough for the purposes of this study.

Comparison with alternative sources

In order to overcome the aforementioned shortcomings, the findings were cross-checked against the following alternative sources: *Ethnologue* (Lewis et al. 2014), the latest available

national censuses, literacy reports, DHS, the documents of the Organisation Internationale de la Francophonie (OIF),[9] Albaugh (2014), and other available documents on individual countries. Data are presented and discussed in the online supplementary material.

The general conclusion that can be derived from the online supplementary material is that the reported size of ethnic and linguistic groups varies considerably across our consulted sources. The discrepancy is the most striking in the case of other than home languages. If estimates are available at all, the share of respondents reporting proficiency in the largest indigenous or the former coloniser's language is regularly the highest in the Afrobarometer and the lowest in the *Ethnologue*.

There are several explanations for the incongruity in the available estimates. To start with, self-reported language proficiency, as already noted, is likely to be biased by the respondent's beliefs on the surveying agency and the purpose of the survey, the interviewer's ethnicity and the social or political status of the language in question.

Second, ethnicity- and language-related variables covered in various surveys are often assumed to refer to the same theoretical concept. In empirical development studies focusing on the socioeconomic impacts of diversity, it is a common practice to identify ethnicity with linguistic data (Cheeseman & Ford 2007). *Ethnologue* (Lewis et al. 2014) makes the same simplification in some cases: information on tribal affiliation from the 2009 Kenyan census (Kenya National Bureau of Statistics 2010) is reported as linguistic data. The 2010 Zambian census (Central Statistical Office 2012), which surveys ethnicity and home language separately, provides evidence that the size of an ethnic group can be remarkably different from the size of the corresponding linguistic group. Bemba and Chewa (or Nyanja) serve as home language for ethnic groups other than the Bemba and the Chewa. However, the Zambian census is unique in this respect; most of the countries do not collect information on both ethnicity and language. The population censuses of Ghana, Liberia, Senegal, and Uganda includes a question on ethnicity only, while Botswana, Burkina Faso, Mali, Mozambique, Namibia, and South Africa survey home languages. Benin applies the sociolinguistic affiliation as the basic classification concept. Kenya and Malawi apply the term 'tribe' in the questionnaire instead of ethnicity. The Nigerian, Tanzanian, and Zimbabwean censuses do not include language- or ethnicity-related questions at all.

An additional source of discrepancy is that the classification schemes applied by population censuses and other surveys follow diverse conceptual principles and, as a consequence, are not equally refined. *Ethnologue*, which works with the highest level of differentiation, usually lists many more groups than any of the remaining sources. Let us consider the case of Benin. Comparing group shares obtained from Afrobarometer, *Ethnologue*, and the 2002 census (INSAE 2003), suggests that respondents in Afrobarometer whose own groups are not listed chose the closest possible one from the predefined list. Thus, for instance, the speakers of Gbe languages are very likely to be included in the Fon group. When we follow the concept of the census and add up the speakers of individual languages belonging to the same broader sociolinguistic group, the reported shares become more comparable across sources.

And finally, the causes of the high differences across sources related to the use of additional languages are discussed. Since they are usually not surveyed directly, the study relies on various materials such as the OIF website (2010), individual estimates collected in Albaugh (2014), and literacy data from the latest population censuses, literacy reports and the DHS to approximate the spread of languages beyond the primary

language level. However, censuses and literacy reports predominantly focus only on literacy in languages in which education is available.

The reported share of the population being proficient in local and European languages is highly dependent on the literacy measurement method. Although the population censuses of some countries (for example Benin, Ghana, Mozambique, Namibia, Senegal) provide information on the respondents' reading and writing skills, Ghana (Ghana Statistical Service 2012) is the only one where these abilities are actually tested. The DHS apply a mixed technique to determine the share of the literate population: individuals with higher than primary education are automatically assumed to be able to read and write, while others are tested if they could read a simple sentence. An additional difficulty that limits the possibility of data collection is that the censuses and the DHS ask if the respondents are literate in any languages, but, reading and writing skills in individual languages, except for those in English, French, and Portuguese, are not presented separately. It is only Botswana and Nigeria that regularly conduct separate countrywide literacy surveys. But, while Botswana measures reading, writing and oral language skills apart, the Nigerian survey is based on self-report.

The next cause of the diversity in literacy data is that the investigated population varies across surveys. The DHS cover citizens aged between 15 and 49. The age threshold below which national censuses do not ask literacy varies between countries.

The strategy of utilising literacy data to gain more insight into the use of languages raises a number of crucial questions. What is meant by language abilities in the different sources? How do reading and writing abilities mirror oral proficiency? As is demonstrated in the 2003 Botswana Literacy Survey (Central Statistics Office 2005), measured writing, reading and communication skills can differ significantly. Whereas 38 per cent and 34.2 per cent of the investigated population had high competence in writing and reading in English respectively, the share of people with high oral competence was only 2.4 per cent (Central Statistics Office 2005, Table 38: 112). But, is it relevant to distinguish between reading, writing and oral skills? It depends on the goal of the study for which the data are used. If linguistic information is used to measure the share of the population that are excluded from political decision-making because the media, documents and the voting-papers are available only in official languages, reading and understanding skills are the most relevant. But, if the study is focused on the efficiency of common action within a community, information on the ability of verbal communication in a certain language could be eligible for the analysis.

Although, due to the aforementioned issues, the reconciliation of the data is difficult, the data found that the shares of people with communication and literacy abilities in the former colonisers' languages are quite similar across sources in about half of the 20 countries (Botswana, Lesotho, Madagascar, Malawi, Mali, Nigeria, Senegal, Tanzania, and Zimbabwe). However, *Ethnologue* usually reports much lower shares compared to Afrobarometer or the literacy information provided in the discussed surveys. In the other half of the countries, with the exception of Benin which is characterised by relatively moderate differences, the study found striking anomalies.

The ICP

The need for a linguistic diversity (also called fragmentation and heterogeneity) measure that accounts for multilingualism has long been recognised in linguistics and political

science. In an early work that systemises the possible approaches, Joseph Greenberg (1956) discusses two types of linguistic indicators that assume monolingual citizens and six other types that handle proficiency in multiple languages. However, partly due to data availability problems, sociolinguistic, development, and political studies investigating the impacts of ethnolinguistic fragmentation (discussed under 'The use of linguistic data' above) are still based on indicators with the limited approach of monolingual citizens.

There are only a few empirical works that reveal the channels through which second languages affect bilingual and multilingual societies. Katalin Buzasi (2015) finds evidence that African people living in regions with higher average ICP (the main measure in this work) are more likely to trust unknown people. Aspachs-Bracons, Clots-Figueras & Masella (2007) show that individuals who experienced more exposure to Catalan language at school after the introduction of the bilingual education system in 1983 were more likely to feel more Catalan than Spanish. What is more, this result persisted among pupils whose parents did not have Catalan origins.

Despite its limitations discussed in the previous section, Afrobarometer provides a unique opportunity to finally elaborate a linguistic measure that accounts for multilingualism, if not at the global level, at least in a number of sub-Saharan African countries. The study applies the ICP (Buzasi 2015),[10] which is based on individual linguistic repertoires obtained from Q3 on home languages and Q88E on additional languages. Due to the data collection and reporting method of the Afrobarometer, the ICP can be computed for individuals and be aggregated at the country (or any desired) level. Technical details on the construction of ICP are provided in Appendix A. The individual ICP scores (Formula A.2) are understood as the probability that one can communicate with a randomly selected other person within the country given one's language repertoire. Country level ICPs (Formula A.3) are computed as the weighted averages of the individual ICPs and can be interpreted as the probability that any two randomly selected people within the society can communicate with each other since they have at least one common language. Although in the above introduced form the ICP captures the linguistic resemblance of citizens rather than their dissimilarity, deducting the ICP from one can be interpreted as the probability that two randomly selected people have no language in common. The ICP is highly correspondent with the concept of the final and most advanced linguistic diversity measure by Greenberg (1956), called the index of communication.

In order to find out how much difference it makes to account for multilingualism in terms of linguistic fragmentation, the simplest forms of ethnic and linguistic heterogeneity measures (Appendix B), which are utilised in development studies, are presented in parallel to the ICP in Table 1. Using Q79 on ethnic affiliation and Q3 on home languages from Afrobarometer, the study computes the probability that two randomly chosen people in a country belong to different ethnic and linguistic groups, respectively. Since ethnicity in the Cape Verdean questionnaire rather refers to social identity and is incomparable with those in other countries, the study does not compute the ethnic diversity measure for this country.

Table 1 reveals two important facts. First, it provides evidence that although ethnic and linguistic fragmentations coincide in the majority of cases, they differ significantly in certain countries. The high gaps between the two heterogeneity measures in Botswana, Lesotho, Madagascar, and Zimbabwe can be explained by the survey design that the dialects of certain languages (Tswana in Botswana, Sotho in Lesotho, Malagasy in

Table 1. Ethnic and linguistic fragmentation, and the ICP in the Afrobarometer countries.

Country (the number of respondents)	Ethnic fragmentation	Linguistic fragmentation	ICP
Benin (1,200)	0.825	0.816	0.581
Botswana (1,200)	0.923	0.407	0.984
Burkina Faso (1,200)	0.688	0.703	0.602
Cape Verde (1,264)	-	0.005	0.995
Ghana (1,200)	0.755	0.718	0.751
Kenya (1,104)	0.890	0.892	0.917
Lesotho (1,200)	0.888	0.040	1.000
Liberia (1,200)	0.888	0.885	0.598
Madagascar (1,350)	0.826	0.020	1.000
Malawi (1,200)	0.781	0.728	0.884
Mali (1,232)	0.839	0.719	0.803
Mozambique (1,200)	0.874	0.872	0.697
Namibia (1,200)	0.705	0.701	0.816
Nigeria (2,324)	0.856	0.876	0.622
Senegal (1,200)	0.701	0.605	0.892
South Africa (2,400)	0.866	0.855	0.606
Tanzania (1,208)	0.954	0.950	0.991
Uganda (2,431)	0.896	0.896	0.484
Zambia (1,200)	0.884	0.872	0.663
Zimbabwe (1,200)	0.827	0.331	0.871
Mean	0.835	0.645	0.788

Note: Since almost everyone speaks Cape Verdean Creole as home language, Q88E on additional languages is not included in the questionnaire of Cape Verde.

Madagascar, and Shona in Zimbabwe) are not distinguished in Q3 on home languages but acknowledged as separate ethnic groups in Q79. However, this issue is at least as theoretical as statistical. Computing a diversity measure based on Q3 is more applicable for studies that focus on communication possibilities provided by common languages and the distinction between sub-groups that easily communicate with each other makes no sense. Or as another option, Q79 and Q3 in the listed countries might be seen as a minimalist and maximalist philosophy to differentiate between groups.

The gap between the two diversity indicators is 12 percentage points (0.839–0.719) in Mali and about 10 percentage points (0.701–0.605) in Senegal. The main explanation for this relatively small but considerable difference is that the largest languages are named as home language by respondents belonging to different ethnic groups. In Mali,

Bambara is mentioned as home language by 44.4 per cent of the Malinke people, by 49.7 per cent of the Peulh/Fulfulde group, by 28.3 per cent of the Senoufo/Mianka ethnic group, and by 25.7 per cent of the Soninke/Sarakolle people. The case of Wolof is similar in Senegal: 35.5 per cent of the Serer, 22.3 per cent of the Pulaar/Toucouleur, and 19.7 per cent of the Mandinka/Bambara reported Wolof as the primary language at home.[11]

The second main conclusion derived from Table 1 is that although societies with low ethnic and linguistic heterogeneity exhibit relatively high average communication potential, ethnically and linguistically highly fragmented countries do not necessarily suffer from poor communication possibilities. Due to the promotion of Swahili as the national language after independence as a crucial part of the nation-building, Tanzania exhibits high average communication potential despite the high levels of ethnic and linguistic fragmentation. A similarly high communication potential can be assigned to Swahili in Kenya, Chewa in Malawi, and Wolof in Senegal. These languages are spoken by more than 90 per cent of the population according to the Afrobarometer. Since Tswana, Sotho, and Malagasy are basically spoken by each respondent, the ICP reaches its maximum value in Lesotho and Madagascar and is above 0.98 in Botswana. Although the linguistic diversity is much smaller than the ethnic diversity in Zimbabwe, the ICP is 'only' 0.871. The reason for this is that a considerable share, 23.91 per cent of the Ndebele speakers who represent more than 10 per cent of the population tends to be monolingual and 43.47 per cent of the multilingual Ndebele people do not speak Shona. The ICP seems to be the lowest in countries where there are regionally dominant languages such as Fon, Adja, Yoruba, and Bariba in Benin; Bemba, Tonga, Chewa, and Tumbuka in Zambia; Hausa, Yoruba and Igbo in Nigeria; and the 11 official languages in South Africa.

A Graphic Representation of the ICP

This section demonstrates how the main dimensions of the linguistic situation in the sample countries can be shown in an insightful way with a simple graphic representation of the ICP. As the first step, the study sorted languages by their size as an additional language in each country and recalculated the ICPs excluding these languages one by one from the database. Languages omitted in a previous step are excluded from the following steps as well. Dark-coloured lines in Figure 1 show the decrease in the communication potential when languages listed on the vertical axis are excluded as additional languages only but still included as home languages. Light-coloured lines show the drop when languages are completely (both as home and additional languages) ignored. In other words, the lines show how high the communication potential would be in a society if listed languages were not spoken as an additional language (dark-coloured) or were not spoken at all (light-coloured). The magnitude of the decrease and the difference between these lines refer to the importance of languages in determining communication potential and reveal some crucial language use patterns. Since additional languages are not surveyed in Cape Verde, this country is not included in Figure 1.

For instance, let us consider the case of Ghana in Figure 1 and in the online supplementary material. The order of languages according to their reported frequency as additional language is English (47.39 per cent), Akan (31.57 per cent), Ga/Dangme (11.04 per cent), and Ewe (4.79 per cent). Although the share of respondents reporting English is above

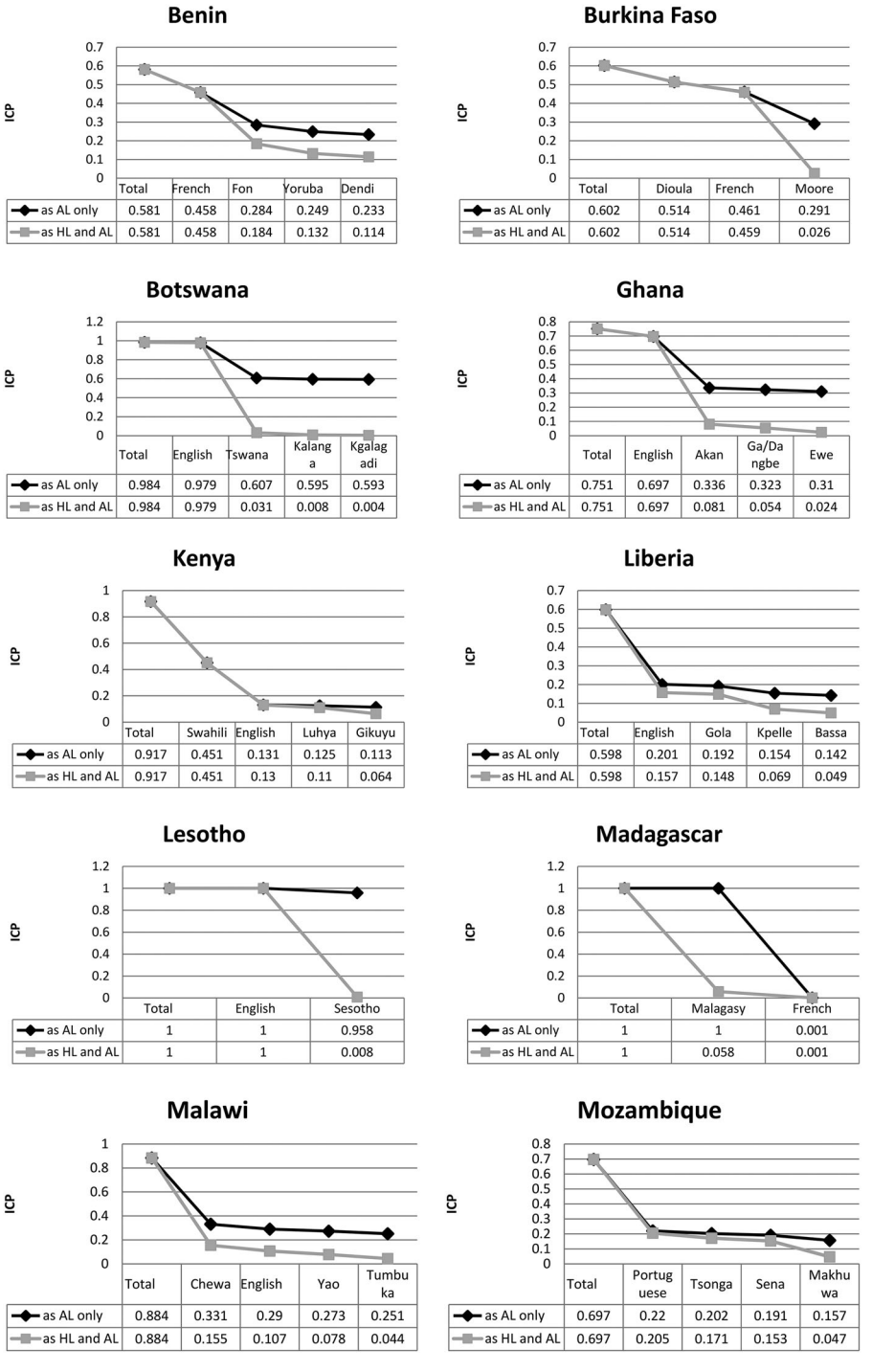

Figure 1. The drop in the communication potential by languages excluded as additional language (AL) only and both as home (HL) and additional language (AL).

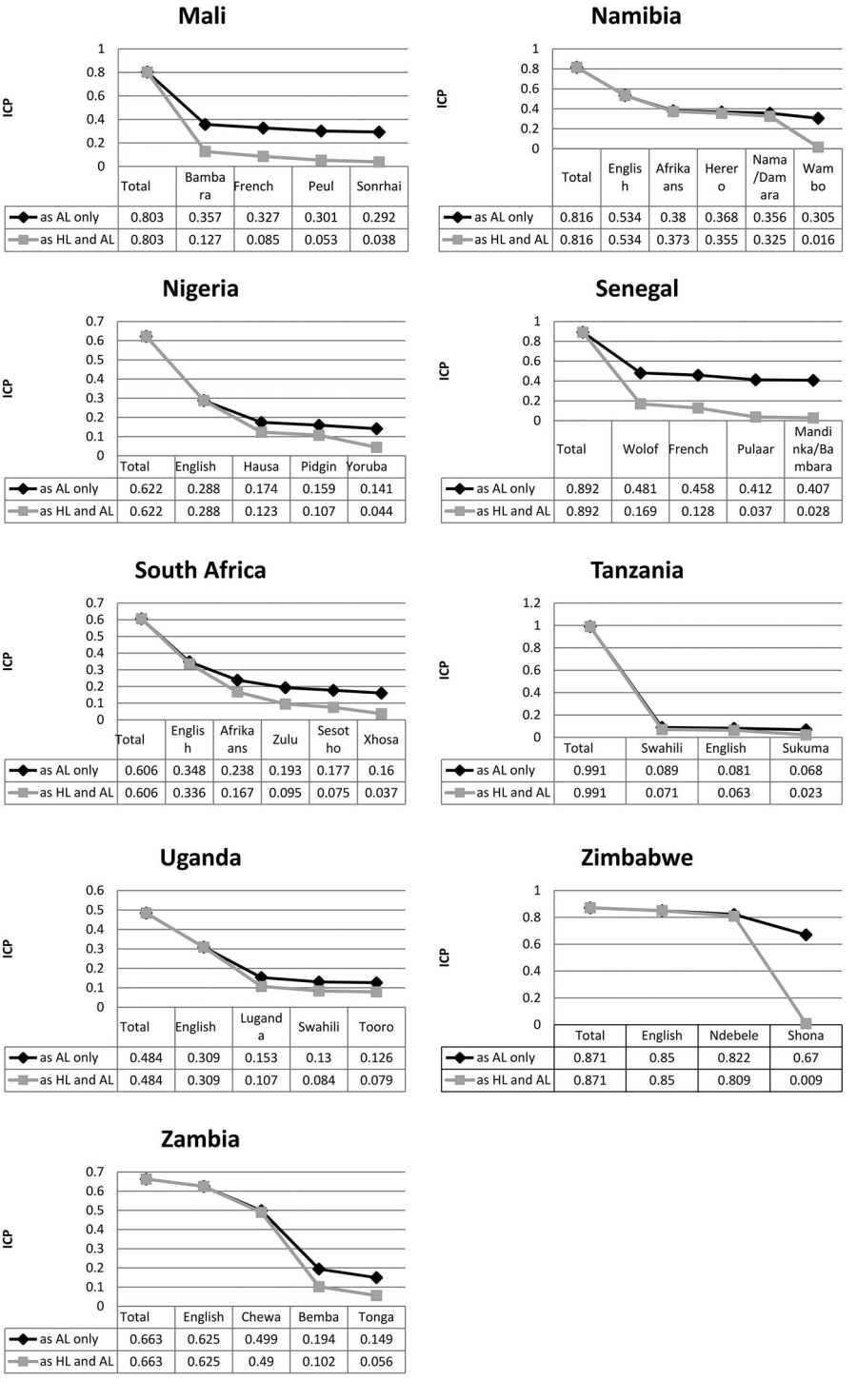

Figure 1. Continued

47 per cent, the drop of the average communication potential when excluding it is only about 5 percentage points. The reason behind this phenomenon is that English is usually not spoken as home language, thus when it is excluded, indigenous languages, spoken either as home or additional language, still 'maintain' the observed level of the communication potential. The exclusion of Akan, the largest indigenous language group in Ghana, contributes to a significant drop. The average communication potential reduces to 0.336 when, in parallel to English, Akan is omitted as an additional language. When Akan is ignored completely, the communication potential decreases even more radically to 0.081. The large gap between the two communication potentials when Akan is excluded as an additional language only and as both home and additional language indicates that people speaking Akan as primary language are very likely to remain monolingual or speak English as the only additional language which has already been omitted in the first step. Table 2 reinforces this argument: 41.44 per cent of the Akan group is found to be monolingual and 28.55 per cent reports English as their only other language. Overall, we find that languages other than English and Akan account for a communication potential of less than 0.1 in Ghana.

At this point, it may be noted that if languages are ordered according to a different aspect such as their size as home language or their total number of speakers, figures highlight other dimensions of the linguistic situation. If Akan was the first language to be omitted and English only afterwards, it would immediately be clear if Akan speakers are more likely to learn English or if they would rather remain monolingual; which phenomenon is less easy to see when English is taken out first.

To make it easier to interpret the graphic representation of the ICP, the article discusses the case of Liberia where the language situation is significantly different from that in Ghana (see Figure 1 and the online supplementary material). The coloniser's language is selected as home language by 23.43 per cent of the sample. None of the local languages is spoken by more than 10 per cent as an additional language. The order of languages according to their reported frequency in Q88E is English (48.96 per cent), Gola (9.61 per cent), Kpelle (8.52 per cent), and Bassa (5.48 per cent). Unlike in Ghana, English plays a significant role in supporting communication potential: when excluding English as an additional language, the ICP drops from 0.598 to 0.201, and when it is completely ignored, the ICP drops to 0.157. The gap between the dark- and light-coloured lines suggests that a significant share of people speaking English the most often at home are

Table 2. The linguistic repertoire of people speaking Akan at home.

L1	L2	share
Akan	-	41.44%
	English	28.55%
	Ga/Dangme	1.31%
	Ewe	0.33%
	English+Ga/Dangme	5.55%
	English+Ewe	2.45%
	Ga/Dangme+Ewe	0.33%
	English+Ga/Dangme+Ewe	0.65%
	Other	19.39%
Total		100%

Note: 4.4% of the Akan speak Nzema, 3.1% speak Hausa, 3.1% speak Sehwi, 2.77% speak French.

monolingual. The drop in ICP when Kpelle is excluded completely is larger than when it is excluded only as an additional language. This can mean that the large share of Kpelle-speaking citizens are monolingual or speak English and/or Gola which have been already taken out in the previous steps. Again, rearranging the language order would reveal which one is the case.

Figure 1 also indicates how many languages the typical citizen speaks. The average number of languages is expected to be the highest in countries where the communication potential decreases moderately when we exclude languages one by one and the gap between the dark- and light-coloured lines remain relatively small at each step. Based on this logic and the shape of the dark and light grey lines, Zambia and Namibia should be ranked as countries with the highest average number of languages. Figure 2, which represents the weighted average of languages in the individual repertoires in each country, reinforces our expectations and ranks these two countries as first and third, respectively. However, countries where Swahili is widely used needs to be discussed separately. In Tanzania and Kenya, where in parallel to home languages Swahili is spoken by almost everyone as an additional language, the average number of mastered languages should be close to two. Since, more than half the population is proficient in English along with home language and Swahili, Kenya is ranked second in Figure 2. Observing Figures 1 and 2, we can arrive at the following rule of thumb: the number of languages in the typical repertoire is above average in countries where the light-coloured line does not drop much below 0.1 after the third language is excluded.

The major benefit of the graphic representation is that it is applicable to analyse the relation between indigenous and European languages. Although in 12 out of the 19 countries the former coloniser's language is reported as the most common additional language in Q88E, the exclusion of English and French from the ICP does not result in a large drop in the majority of the sample countries (Burkina Faso, Ghana, Madagascar, Malawi, Mali, Senegal, Tanzania, Zambia, and Zimbabwe). English and French contribute effectively to the communication potential only in a few cases: without English, the communication potential would be 0.157 instead of 0.598 in Liberia, 0.534 instead of 0.816 in Namibia, 0.288 instead of 0.622 in Nigeria, 0.348 instead of 0.606 in South Africa, and 0.309

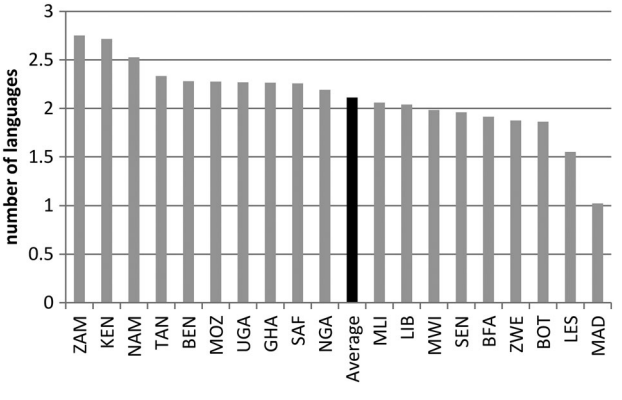

Figure 2. The number of languages in the typical repertoire per country.
Note: Figure 2 presents weighted average. Sample weights are obtained from Afrobarometer.

instead of 0.484 in Uganda. Swahili and English account for almost all communication possibilities in Kenya. The role of Portuguese in Mozambique is similar to that of English in the above discussed countries: when we ignore Portuguese as an additional language only, the communication potential reduces to 0.22 from 0.697 and to 0.205 when it is completely omitted. Among the five former French colonies, it is only Benin where French seems to determine communication potential significantly.

The most common vernacular is Dioula in Burkina Faso, Chewa in Malawi, Bambara in Mali, and Swahili in Kenya and Tanzania. Swahili and Dioula are similar in the sense that despite their relative low use as a home language they are widely spoken. The large gaps between the dark- and light-coloured lines in Malawi, Mali, and Senegal when the most frequently reported languages are taken out of the sample suggest that the speakers of these linguistic groups are very likely to remain monolingual. Afrikaans contributes to a high proportion of the communication potential in South Africa and Namibia, even though it is spoken by only about 8 per cent as home language in the latter country.

The online supplementary material and the graphic representations classify countries by their language use patterns. According to the distribution of the major indigenous languages, the sample countries can be organised into five groups: the first set of countries consists of Cape Verde, Lesotho and Madagascar where a single indigenous language is spoken by almost all citizens as home language; in the second cluster, Botswana, Ghana, Malawi, Mali, Senegal, and Zimbabwe, the largest indigenous language is spoken by between 80 and 100 per cent of the population but by only about 50 to 80 per cent as home language; the main characteristic of the third group, consisting of Benin and Uganda, is that the largest indigenous language is the most popular vernacular and even though the share of the total speakers does not exceed 50 to 60 per cent, there are not any serious indigenous competitors; in the fourth group, the largest home language is not the most widely spoken as an additional language by other groups – Dioula outnumbers Moore in Burkina Faso and Afrikaans outnumbers Wambo in Namibia as a vernacular, Swahili is spoken by almost everyone in Kenya and Tanzania; the final group, which includes Liberia, Mozambique, Nigeria, South Africa, and Zambia, is characterised by a few relatively large regionally dominant languages.

Adding the languages of the former colonisers to the patterns explained above, the picture on the linguistic situation becomes even more sophisticated. As Figures 3 and 4 present, the distribution of European languages is dependent on the distribution of vernaculars. The share of the population proficient in the former coloniser's language is negatively associated with the size of the largest home language (Figure 3) and the size of the most widely spread language (Figure 4). The latter two groups of countries in the classification scheme introduced above are scattered roughly in the upper left part of Figures 3 and 4. These countries are also the ones where the exclusion of the former coloniser's language results in a considerable drop in the average communication potential in Figure 1. Thus, the language of the former coloniser is the most widely spoken and the most important in terms of the communication potential in countries where there is no indigenous language that serves as a national lingua franca. However, without undermining the validity of this general pattern, there might be differences between countries formerly colonised by different nations. While English and the most widely spread vernacular seem to be complementaries in former British colonies, in Burkina Faso, Madagascar, Mali, and Senegal, where a local alternative is available, proficiency in French remains relatively

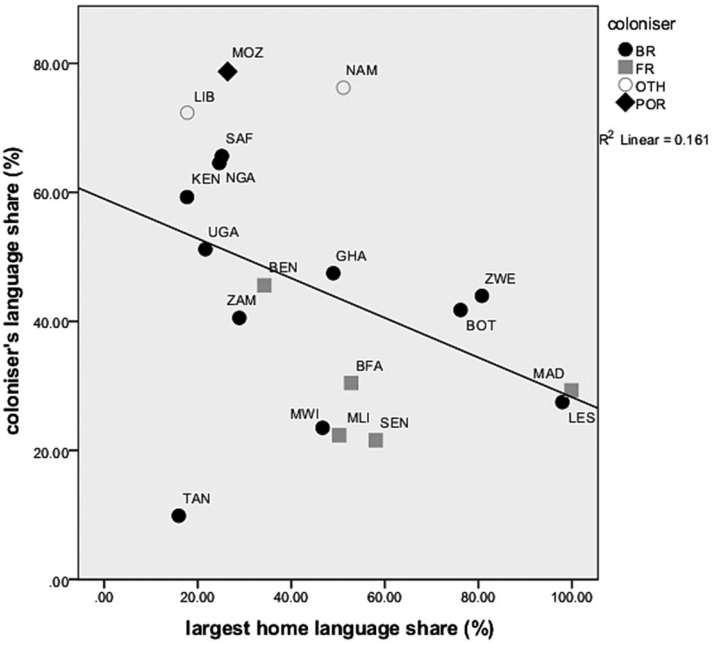

Figure 3. The relationship between the size of the largest home language and the share of the population speaking the former coloniser's language.

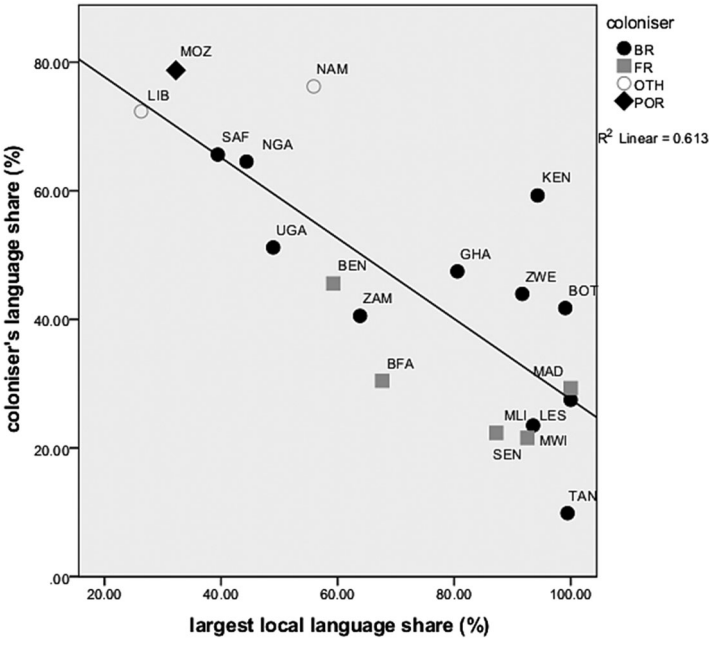

Figure 4. The relationship between the size of the most widely spoken and the former coloniser's language.

low. The situation in Benin fits more into the general pattern and is similar to those in other than French colonies: the relatively small largest local language is accompanied by a relatively high French proficiency. If Niger and Guinea, which would be located on the left side of the horizontal axis in Figure 3, were included in the Afrobarometer, there would be a better chance to find out if the low French proficiency and the preference for a local language is a general pattern in former French colonies or is likely to be a special attribution of countries located on the right side of the horizontal axis in Figure 3. Moreover, the findings suggest that the exclusive use of French in education and administration in former colonies does not necessarily lead to the weakening of local languages, and the recognition of indigenous languages in former British colonies does not reduce the demand for English.

Conclusion and Findings

Utilising Round 4 f the Afrobarometer Survey, this study presents the most important dimensions of the linguistic situation in 20 sub-Saharan African countries. Without repeating what has already been discussed in the previous sections, the conclusion is devoted to illustrating the relevance of the findings for policy-makers and social and political scientists engaged in language-related issues.

Development researchers, economists and political scientists are most interested in the potential negative societal impacts of ethnic and linguistic diversity. Based on the discussion in the sub-section titled 'Comparison with alternative sources', this study argues that, even if ethnic or linguistic diversity is computed by a certain formula, the calculated values are likely to be dependent on the design of the underlying material from which the data are retrieved. By comparing the Afrobarometer to alternative sources, the article identifies five survey design-related factors that influence the observed linguistic situation: (1) the detailedness of the classification scheme in the questionnaire; (2) the data collection method; (3) the properties of the investigated population; (4) the purpose and the conceptual framework of the survey; and (5) the respondents' behaviour. The article suggests that these five factors should be kept in mind when the severity of diversity within a country is investigated or when two societies are compared based on various sources. In addition, these findings are expected to be helpful in designing surveys that involve ethnicity- and language-related questions.

The article also indicates that taking other than first and home languages into account makes it possible to analyse some aspects of the linguistic situation that have gained only marginal attention so far. Since, as Table 1 suggests, a society's communication potential is not necessarily determined by its ethnic or linguistic heterogeneity, the investigation of multilingualism, a societal characteristic that potentially counterbalances the harmful effects of diversity, is a promising new direction in development and political research. And lastly, the article suggests that the graphic representation of the ICP is easily adjustable for various research goals such as the classification of countries according their language use patterns. While the main text avoided language policy evaluation or language planning suggestions, it is easy to see that if suitable data on individual language repertoires are available, the ICP can be applied to evaluate the efficiency of language-related programmes and to monitor language dynamics.

Notes

1. Afrobarometer Data (Benin, Botswana, Burkina Faso, Cape Verde, Ghana, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mozambique, Namibia, Nigeria, Senegal, South Africa, Tanzania, Uganda, Zambia, Zimbabwe) (Round 4 2008, 2009) <http://www.afrobarometer.org> (accessed 26 January 2015).
2. See WALS Project <http://wals.info/>; OLAC <http://www.language-archives.org/> (accessed 23 March 2015).
3. See <http://joshuaproject.net> (accessed 23 March 2015).
4. See Polomé (1982) for a detailed overview.
5. See SIL International <http://www.sil.org>; UNESCO Endangered Languages Project <http://www.unesco.org/new/en/culture/themes/endangered-languages/>; Catalogue of Endangered Languages <http://www.endangeredlanguages.com>; Moro Language Project financed by the National Science Foundation <http://moro.ucsd.edu> (accessed 23 March 2015).
6. Round 1 (12 countries, 1999–2001); Round 2 (16 countries, 2002–2004); Round 3 (18 countries, 2005–2006); Round 4 (20 countries, 2008–2009); Round 5 covering 36 countries, including those in northern Africa, was being processed and digitalised in 2015; Round 6 is under preparation.
7. The goal is to give every adult citizen an equal and known chance of selection for the interview. This is achieved via (1) using random selection methods at every stage of sampling; and (2) sampling at all stages with probability proportionate to population size wherever possible to ensure that larger (i.e. more populated) geographic units have a proportionally greater probability of being chosen for the sample.
8. Benin, Botswana, Burkina Faso, Cape Verde, Ghana, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mozambique, Namibia, Nigeria, Senegal, South Africa, Tanzania, Uganda, Zambia, and Zimbabwe.
9. See <www.francophonie.org/>.
10. The index is different from the Q-value of communication potential introduced by Abram de Swaan (1993). Although both indicators attempt to measure the value of language repertoires in terms of the share of a population that can be reached, their main aim and construction are different. Originally, the Q-value is designed to show the communication potential of language repertoires in the European Union and its change in time due to the admission of new member states. Later, the Q-value of certain languages and repertoires was computed for Congo/Zaire (De Swaan 1996), Senegal, and South Africa (De Swaan 2001).
11. Groups are named and spelled as in the codebooks of Round 4 of the Afrobarometer Survey.

Acknowledgements

The author is grateful to Maarten Mous (Leiden University), Peter Foldvari (Utrecht University), the two anonymous referees for *African Studies*, and the participants of the Second Lisbon Meeting on Institutions and Political Economy (2013) for their invaluable comments on the ICP and the previous versions of this article.

Note on Contributor

Katalin Buzasi is a researcher at the Faculty of Humanities at Utrecht University and lecturer in the PPLE College at the University of Amsterdam. Her research interests include the economics of language, diversity measurement in multilingual societies, the link between languages and socioeconomic development, the deep roots of language survival and decline, and the possibilities of analysing language policy and planning with the tools of economics in sub-Saharan Africa.

References

Academija nauk SSSR. 1964. *Atlas Narodov Mira*. Moscow: Glavnoe Upravlenie geodezii i kartografii.

Albaugh, E.A. 2014. *State-building and Multilingual Education in Africa*. Cambridge: Cambridge University Press.

Alderson, J.C. 2005. *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*. London: A&C Black.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S. & Wacziarg, R. 2003. 'Fractionalization'. *Journal of Economic Growth* 8(2): 155–94.

Alesina, A. & La Ferrara, E. 2002. 'Who trusts others?' *Journal of Public Economics* 85(2): 207–34.

Asher, R.E. & Moseley, C. (eds). 2007. *Atlas of the World's Languages*. London: Routledge.

Aspachs-Bracons O., Clots-Figueras, I. & Masella, P. 2007. 'Identity and language policies'. Universidad Carlos III de Madrid Working Paper (Economics) 07–46 <http://e-archivo.uc3m.es/handle/10016/2363> (accessed 26 January 2015).

Baetens Breadsmore, H. 1982. *Bilingualism: Basic Principles*. Clevedon: Tieto Ltd.

Baldauf, R.B. & Kaplan, R.B. (eds). 2004. *Language Policy and Planning in Africa, Vol. 1, Botswana, Malawi, Mozambique and South Africa*. Clevedon: Multilingual Matters Ltd.

Banton, M. 1997. *Ethnic and Racial Consciousness 2nd edition*. New York: Longman.

Batibo, H.M. 2005. *Language Decline and Death in Africa. Causes, Consequences and Challenges*. Clevedon: Multilingual Matters Ltd.

Brown, G.K. & Langer, A. 2010. 'Conceptualizing and measuring ethnicity'. *Oxford Development Studies* 38(4): 411–36.

Burton, J., Nandi, A. & Platt, L. 2010. 'Measuring ethnicity: Challenges and opportunities for survey research'. *Ethnic and Racial Studies* 33(8): 1332–49.

Buzasi, K. 2015. 'Languages, communication potential and generalized trust in sub-Saharan Africa: Evidence based on the Afrobarometer Survey'. *Social Science Research* 49(1): 141–55.

Capo, H.B., Gbeto, F. & Huannou, A. (eds). 2009. *Langues Africaines dans l'enseignement au Benin: Problemes et Perspectives*. Cape Town: CASAS.

Central Statistics Office. 2005. Report of the Second National Survey on Literacy in Botswana. Gaborone: The Department of Printing and Publishing Services <http://www.cso.gov.bw/templates/cso/file/File/literacy_report03.pdf> (accessed 26 January 2015).

Central Statistical Office. 2012. *2010 Census of Population and Housing. Volume 11: National Descriptive Tables*. Lusaka: Central Statistical Office <http://catalog.ihsn.org/index.php/catalog/4124> (accessed 26 January 2015).

Cheeseman, N. & Ford, R. 2007. 'Ethnicity as a political cleavage'. *Afrobarometer Working Paper* 83 <http://www.afrobarometer.org/files/documents/working_papers/AfropaperNo83.pdf> (accessed 26 January 2015).

Chuo Kikuu cha Dar es Salaam. 2009. *Atlasi ya lugha za Tanzania*. Dar es Salaam: Chuo Kikuu cha Dar es Salaam.

De Swaan, A. 1993. 'The evolving European language system: A theory of communication potential and language competition'. *International Political Science Review* 14(3): 241–55.

De Swaan, A. 1996. 'La francophonie en Afrique: Une vision de la sociologie et de l'economie politique de la langue', in J.-L. Calvet & C. Juillard (eds), *Les Politiques Linguistiques, Mythes et Réalités*. Montreal: AUPELF-UREF.

De Swaan, A. 2001. *World of the Words: The Global Language System*. Cambridge: Polity Press.

Desmet, K., Ortuno-Ortin, I. & Wacziarg, R. Forthcoming. 'Linguistic cleavages and economic development', in V. Ginsburgh & S. Weber (eds), *Palgrave Handbook of Economics and Language*. London: Macmillan.

Easterly, W. & Levine, R. 1997. 'Africa's growth tragedy: Policies and ethnic divisions'. *The Quarterly Journal of Economics* 112(4): 1203–50.

Eifert, B., Miguel, E. & Posner, D.N. 2010. 'Political competition and ethnic identification in Africa'. *American Journal of Political Science* 54(2): 494–510.

Fearon, J.D. 2003. 'Ethnic and cultural diversity by country'. *Journal of Economic Growth* 8(2): 195–222.

Fishman, J.A. 1991. *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Clevedon: Multilingual Matters.
Gass, S.M. & Selinker, L. 2008. *Second Language Acquisition: An Introductory Course 3rd edition*. New York & London: Routledge.
Ghana Statistical Service. 2012. *2010 Population and Housing Census. Summary Report of Final Results*. Accra: Sakoa Press Limited <http://www.statsghana.gov.gh/docfiles/2010phc/Census2010_Summary_report_of_final_results.pdf> (accessed 26 January 2015).
Greenberg, J.H. 1956. 'The measurement of linguistic diversity'. *Language* 32(1): 109–15.
Hale, H.E. 2004. 'Explaining ethnicity'. *Comparative Political Studies* 37(4): 458–85.
Harbert, W., McConnell-Ginet, S., Miller, A. & Whitman, J. 2009. *Language and Poverty*. Clevedon: Multilingual Matters.
Heine, B. & Möhlig, W. J. G. 1980. (eds). *Language and Dialect Atlas of Kenya, vol 1, Geographical and Historical Introduction: Language and Society, Selected Bibliography*. West-Berlin: Dietrich Reimer.
Herfindahl, O.C. 1950. 'Concentration in the steel industry'. PhD thesis, Columbia University.
Horowitz, D.L. 1985. *Ethnic Groups in Conflict*. Berkeley: University of California Press.
INSAE. 2003. *Troisieme Recensement General de la Population et de l'Habitation. Synthese des analyses en bref*. Cotonou: Direction des Etudes Demographiques <http://www.insae-bj.org/recensement-population.html> (accessed 26 January 2015).
Kaplan, R.B. & Baldauf, R.B. (eds). 2007. *Planning and Policy in Africa. Vol. 2. Algeria, Côte d'Ivoire, Nigeria and Tunisia*. Clevedon: Multilingual Matters.
Kenya National Bureau of Statistics. 2010. *The 2009 Kenya Population and Housing Census. Volume II: Population and Household Distribution by Socio-Economic Characteristics* <http://www.knbs.or.ke/index.php?option=com_phocadownload&view=category&id=109:population-and-housing-census-2009&Itemid=599> (accessed 26 January 2015).
Laitin, D.D. 2000. 'What is a language community?' *American Journal of Political Science* 44(1): 142–55.
Laitin, D.D. 2007. *Language Repertoire and State Construction in Africa*. Cambridge: Cambridge University Press.
Lewis, M.P. (ed). 2009. *Ethnologue: Languages of the World 16th edition*. Dallas: SIL International <http://archive.ethnologue.com/16/> (accessed 26 January 2015).
Lewis, M.P. & Simons, G.F. 2010. 'Assessing endangerment: Expanding Fishman's GIDS'. *Revue Roumaine de Linguistique* 55(2): 103–20.
Lewis, M.P., Simons, G.F. & Fennig, C.D. (eds). 2014. *Ethnologue: Languages of the World 17th edition*. Dallas: SIL International <http://www.ethnologue.com> (accessed 26 January 2015).
MacIntyre, P.D., Noels, K.A. & Clément, R. 1997. 'Biases in self-ratings of second language proficiency: The role of language anxiety'. *Language Learning* 47(2): 265–87.
Maho, J.F. 1998. *Few People, Many Tongues: The Languages of Namibia*. Windhoek: Gamsberg Macmillan.
Mauro, P. 1995. 'Corruption and growth'. *The Quarterly Journal of Economics* 110(3): 681–712.
Mesthrie, R., Swann, J., Deumert, A. & Leap, W.L. 2009. *Introducing Sociolinguistics 2nd edition*. Edinburgh: Edinburgh University Press.
Montalvo, J.G. & Reynal-Querol, M. 2005. 'Ethnic polarization, potential conflict and civil wars'. *American Economic Review* 95(3): 796–816.
Montalvo, J.G. & Reynal-Querol, M. 2010. 'Ethnic polarization and the duration of civil wars'. *Economics of Governance* 11(2): 123–43.
Moseley, C. (ed). 2010. *Atlas of the World's Languages in Danger 3rd edition*. Paris: UNESCO <http://www.unesco.org/culture/en/endangeredlanguages/atlas> (accessed 18 April 2015).
Muzale, H.R.T. & Rugemalira, J.M. 2008. 'Researching and documenting the languages of Tanzania'. *Language Documentation and Conservation* 2(1): 68–108.
Namyolo, S. & Nakayiza, J. 2014. 'Dilemmas in implementing language rights in multilingual Uganda'. *Current Issues in Language Planning* 16(4): 409–424.
Nettle, D. & Romaine, S. 2000. *Vanishing Voices. The Extinction of the World's Languages*. Oxford: Oxford University Press.
North, B. 2000. *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang.

Nunn, N. 2010. 'Religious conversion in Colonial Africa'. *American Economic Review Papers and Proceedings* 100(2): 147–52.
Ortega, L. 2009. *Understanding Second Language Acquisition*. Abingdon & New York: Routledge.
Polomé, E.C. 1982. 'Sociolinguistically oriented language surveys: Reflections on the survey of language use and language teaching in Eastern Africa' Review Article. *Language in Society* 11 (2): 265–83.
Pool, J. 1972. 'National development and language diversity', in J.A. Fishman (ed), *Advances in the Sociology of Language, Volume II*. The Hague: Mouton.
Pray, L. 2005. 'How well do commonly used language instruments measure English oral-language proficiency?' *Bilingual Research Journal: The Journal of the National Association for Bilingual Education* 29(2): 387–409.
Putnam, R. 2007. 'E pluribus unum: Diversity and community in the twenty-first century'. *The 2006 Johan Skytte Prize Lecture. Scandinavian Political Studies* 30(2): 137–74.
Rabenoro, M. (ed). 2013. *Langue et Education: Quelle Langue Utiliser en Classe, a Madagascar au 21eme siecle*. Cape Town: CASAS.
Traill, A. 1985. *Phonetic and Phonological Studies of !Xoo Bushman*. Hamburg: Helmut Buske Verlag.
UNESCO. 2000. *World Language Survey: Official Languages of South Africa*. UNESCO: Department of Arts and Culture.
UNESCO. 2003. 'Language vitality and endangerment'. International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages.
Van der Merwe, I.J. & Van der Merwe, J.H. 2006. *Linguistic Atlas of South Africa. Language in Space and Time*. Stellenbosch: Department of Geography and Environmental Studies, Stellenbosch University.

Appendix A

The construction of the ICP

The basis of the individual and country level ICPs is a *nxn* symmetric matrix $\mathbf{M_k}$ (Formula (A.1)) with elements m_{ijk}, where *i* and j refer to individual *i* and *j* (*i* and $j = 1$ to n_k) in country *k* ($k = 1$ to *20*). n_k is the number of respondents in country *k*. If individual *i* is able to communicate with individual *j* in country *k* given their language repertoires, m_{ijk} is 1, otherwise 0. Matrix $\mathbf{M_k}$ is symmetric in the sense that other factors than languages that possibly influence communication between citizens (geographical or linguistic distance, willingness to communicate, and ethnic disinclination) are not taken into account. Moreover, the number of common languages is also ignored.

$$\mathrm{M}_{\mathbf{k}} = \begin{bmatrix} & 1 & 2 & \cdots & j & \cdots & n_k \\ 1 & m_{11k} & m_{12k} & \cdots & m_{1jk} & \cdots & m_{1n_kk} \\ 2 & m_{21k} & m_{22k} & \cdots & m_{2jk} & \cdots & m_{2n_kk} \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots \\ i & m_{i1k} & m_{i2k} & \cdots & m_{ijk} & \cdots & m_{in_kk} \\ \vdots & \vdots & \vdots & & \vdots & \ddots & \vdots \\ n_k & m_{n_k1k} & m_{n_k2k} & \cdots & m_{n_kjk} & \cdots & m_{n_kn_kk} \end{bmatrix} \tag{A.1}$$

The communication potential of individual *i* in country *k* is computed as shown in Formula (A.2).

$$icp_{ik} = \frac{\sum_{j=1, j\neq i}^{n_k} w_{jk} m_{ijk}}{\sum_{j=1}^{n_k} w_{jk} - w_{ik}} = \frac{\sum_{j=1, j\neq i}^{n_k} w_{jk} m_{ijk}}{(n_k - w_{ik})} \tag{A.2}$$

where w_{ik} and w_{jk} are the sample weights for individual *i* and j respectively in country *k* provided by Afrobarometer. Excluding w_{ik} from the numerator and denominator is a necessary correction not to take one's communication potential with oneself into account. The ICP can be interpreted as the

likelihood that individual *i* can communicate with a randomly selected other citizen, *j*, in country *k* given one's language repertoire. Country level ICPs (Formula (A.3)) are computed as the weighted averages of the individual indices and can be understood as the probability that two randomly selected individuals in country *k* can communicate with each other based on common languages.

$$ICP_k = \frac{\sum_{i=1}^{n_k} w_{ik} icp_{ik}}{n_k} \tag{A.3}$$

Appendix B

Ethnic and linguistic diversity

Ethnic and linguistic diversity D_k in country *k* is computed using Formula (4).

$$D_k = 1 - \sum_{g=1}^{G_k} s_{gk}^2 \tag{A.4}$$

where s_{gk} is the share of ethnic or linguistic group *g* in country *k* and G_k is the total number of ethnic or linguistic groups in country *k* obtained from Q79 on ethnicities and Q3 on home languages in Afrobarometer. Formula (4) is also known as one minus the Herfindahl-index of concentration (Herfindahl 1950).