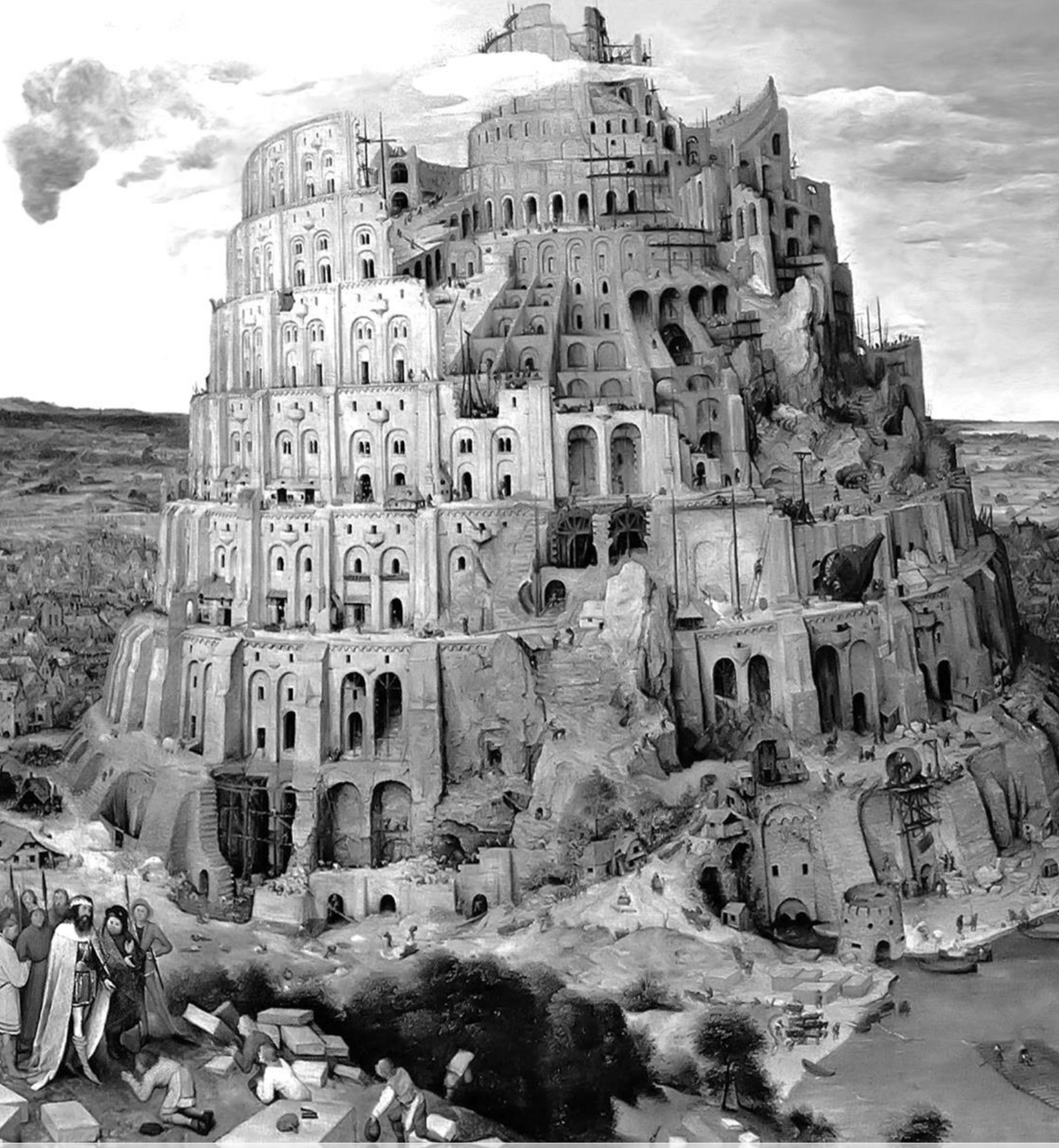


The language situation in Sub-Saharan Africa: Historical roots, measurement, and development impacts

Katalin Buzási

PhD thesis

Cover: The tower of Babel painted by Pieter Breughel the Elder c. 1563 (museum: Kunsthistorisches Museum, Vienna, Austria)

THE LANGUAGE SITUATION IN SUB-SAHARAN AFRICA: HISTORICAL ROOTS, MEASUREMENT, AND DEVELOPMENT IMPACTS

De taalsituatie in Sub-Saharisch Afrika: historische wortels, meettechnieken en ontwikkelingseffecten
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op donderdag 19 november 2015 des ochtends te 10.30 uur

door

Katalin Buzási
geboren op 29 november 1985
te Mátészalka, Hongarije

Promotoren: Prof. dr. J.L. van Zanden

Prof. dr. M.P.G.M. Mous

Acknowledgements

I would like to take the opportunity to express my gratitude to a number of people that have made the successful completion of my PhD project possible. First of all, I wish to thank my promotor, Jan Luiten van Zanden. Without his support I could never have the chance to obtain my PhD degree at a well-established Western European university. Second, I am greatly indebted to my co-promotor, Maarten Mous, who has always encouraged me, provided insightful comments on my work and has always had the time to help me with practical issues as well.

Furthermore, I would like to thank my colleagues at the Department of History and Art History and the Center for Global Economic History for the inspiring working environment. I would like to thank the following persons individually for the friendly atmosphere and the precious moments we have experienced together: Winny Bierman, Sarah Carmichael, Selin Dilli, Miguel Laborda Peman, Mikolaj Malinowski, Felix Meier Zu Selhausen, Kostadis Papaioannou, Sandra de Pleijt, Auke Rijpma, Danielle Teeuwen, Annelies Tukker, Lotte van der Vleuten, and Pim de Zwart. I would like to thank Danielle also for translating the 'Samenvatting' and the abstracts.

I am also grateful to my friends in the Netherlands and Hungary for being supportive and always interested in my progress: Martijn van Deel, Ewout ten Heuw, Leonard van Egmond, Clary Jelsma, Wilma Kaptein, Corine van Middelkoop, Wessel Nordeman, Harma Woldhuis, Maya Wester, Bas van Leeuwen, Jieli van Leeuwen-Li, Diána Balkay, Zsuzsanna Bényei, Béla Budai, Hella Debrecenyi, Bence Egri, Nóra Elek, Anett Flaskár, Judit Futó, Dóra Kerti, Marietta Kiss, István Kovács, Eszter Mózes, Levente Nádasi, Andrea Szabó, Enikő E. Szilágyi, and Máté Vona.

Also, I would like to thank my family, especially my parents, László Buzási and Lászlóné Buzási, my grandma, Tivadarné Katona, for their invaluable support through the ups and downs of my PhD journey (and, of course, before that). And last but not least, I am grateful to Péter Földvári for his personal support, understanding in stressful periods, and constant encouragement in everyday life and in professional issues.

Katalin Buzási
Utrecht, 19 November 2015

Contents

Tables

Figures

1 Introduction

1.1 Motivation

Since the groundbreaking work of North (1990, 1991), institutions have become a central theme in the research on economic and political development. Good quality institutions are commonly referred to as norms, rules or behavior patterns that lower the transaction costs of interactions at the micro level contributing to the efficiency of social sub-systems at the macro level (North 1992). Empirical studies provide evidence that efficiently working institutions such as democracy, the rule of law, the protection of property rights, contract enforcing arrangements, and certain cultural traits are important sources of human development and economic growth (Alchian and Demsetz 1973, Barro 1996, Asoni 2008, Tabellini 2010, Haggard and Tiede 2011).

The inevitable consequence of recognizing the importance of institutions in social sciences has been the increasing reliance on history in the research into the roots of development. Since Robert W. Fogel and Douglass C. North were awarded the Nobel Memorial Prize in Economic Sciences in 1993 for 'having renewed research in economic history by applying economic theory and quantitative methods in order to explain economic and institutional change'[1], an increasing number of articles incorporating historical approach have been published in leading economic journals.

It is quite surprising, though, that while the spectrum of recognized and investigated institutions is almost infinite (Voigt 2013) and the importance of history in determining development in the long-run is widely acknowledged (Acemoglu et al. 2001, Nunn 2007, Huillery 2009), language, one of the oldest and most naturally evolving human institutions, draws only marginal attention. This is so, even though, language is often compared to money regarding its importance and functions (Coulmas 1992). As a medium of exchange, language, through communication, facilitates information flow between people or groups and consequently lowers the costs of cooperative behavior (Smith 2010). As a store of value, it is an essential tool to preserve culture and traditions and transfer those from one generation to the next (Salzmann et al. 2014).

In economics, linguistic issues appear in two ways. First, in the general discourse language is discussed only as an unavoidable side-topic of certain more researched areas. Microeconomic theory, which highlights the importance of communication in reducing transaction costs, views language simply as a communication tool (Ostrom 2000). Similarly, empirical models that explain bilateral trade flows consider the lack of common languages as a factor that increases trading costs (Egger 2002, Martinez-Zarzoso 2003). Probably the most important property of languages from the aspect of

[1] Accessed from: http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1993/ [25 Feb 2015]

macroeconomics and development studies is that they can serve as a basis for social organization (discussed in more detail in Section 1.3.1 and 1.3.2). High ethnolinguistic fragmentation is found to be an important determinant of the underdevelopment of African and Asian societies (Alesina et al. 2003, Putnam 2007, Easterly and Levine 1997). Second, language stands in the focus of a narrow interdisciplinary field within economics entitled the 'economics of language' or 'language(s) and economy' (Zhang and Grenier 2013, Grin 1996). Although this area addresses economic questions directly related to languages, its geographical focus is quite limited to the Western world, namely Canada, Switzerland, Belgium, and the European Union (Vaillancourt 1996, Crystal 1998, Vaillancourt and Coche 2009, van Parijs 2011, Gazzola 2006, Ginsburgh et al. 2005).

The aim of this thesis is to fill the aforementioned gaps in the literature to some extent. Languages are considered as fundamental institutions that are shaped gradually over the long run and have crucial development impacts. Unlike the field of language economics, we concentrate on a less developed region of the world, notably Sub-Saharan Africa, where linguistic issues are different from those experienced in the modern Western countries. The developed world consists of nation states where people, including minority groups, are likely to be proficient in the official language(s) and the linguistic problems are predominantly related to increasing immigration and the endangerment of small local languages. Sub-Saharan African societies are multilingual where the official languages are usually 'imported' from abroad and are mastered by only a relatively small proportion of the population. Indigenous languages are regionally concentrated and officially not recognized. The lack of a single national language is associated with difficulties in nation-building, social cohesion and human development. Although economics and development studies are concerned with the societal effects of ethnolinguistic fragmentation (discussed in Section 1.2.1 and 1.3.1), other dimensions of the language situation such as multilingualism (second languages) or the proficiency in the former colonizer's language is less researched in this respect. It is sociolinguistics and anthropology which are concerned with these issues in less developed societies, but studies in these fields are usually descriptive and focus on one particular linguistic group, region or country. This thesis builds up on the findings of sociolinguistics and anthropology which go beyond the study of ethnolinguistic fragmentation, but aims to arrive at generalizable instead of region- or language-specific results by relying on economic approach, large datasets and modern statistical and econometric techniques.

1.2 The geographical scope of the study: Sub-Saharan Africa

Due to their similar economic and cultural features and common historical roots, the 'Middle East and North Africa (MENA)' is usually treated separately from Sub-Saharan Africa by international organizations (World Bank and the OECD). Since the language patterns of North and Sub-Saharan Africa are substantially different, this thesis takes

over this distinction and does not include the Northern part of the continent. More explanation is provided in Section 1.2.2.

1.2.1 The peculiarities of the underdevelopment of Sub-Saharan Africa

Sub-Saharan Africa (SSA) is commonly acknowledged as the most underdeveloped area of the world. According to the latest GDP per capita data (in current US dollar) (World Bank 2013), the fifteen poorest countries are all located in this region. Moreover, there are only fourteen non-SSA countries among the poorest fifty. Figure 1.1 shows the distribution of GDP per capita per world region in 2013. Although there are some countries which seem to perform relatively well in economic terms, the median GDP per capita is still the lowest in Sub-Saharan Africa (1045 USD). The relatively high values can be attributed to tourism in the Seychelles (SYC), agriculture, services and tourism in Mauritius (MUS), and to oil revenues in Angola (AGO), Equatorial Guinea (GNQ), and Gabon (GAB).

Following Barro (1991), who found that the obvious factors such as human capital, government consumption, market distortions, and political instability cannot explain why African (and Latin-American) economies grew so slowly after 1960, many scholars have attempted to solve the 'mystery' of 'Africa's growth tragedy' (Easterly and Levine 1997).[2] The majority of the literature suggests that it is bad policy choices made by African leaders after independence that are responsible for the observed poor performance. Mauro (1995) finds that corruption, a severe problem in Africa, discourages investments which are essential for economic growth. Sachs and Warner (1997) identify the limited openness to international trade as the main problem. Easterly and Levine (1997) blame the insufficiency of public good provision.

But why did policy makers choose such disadvantageous policies? And why did not policy actions, which worked well in other parts of the world, result in the desired outcomes? By now, it is established that these questions cannot be answered without the understanding of Africa's history and its influence on the development of institutions and social arrangements.

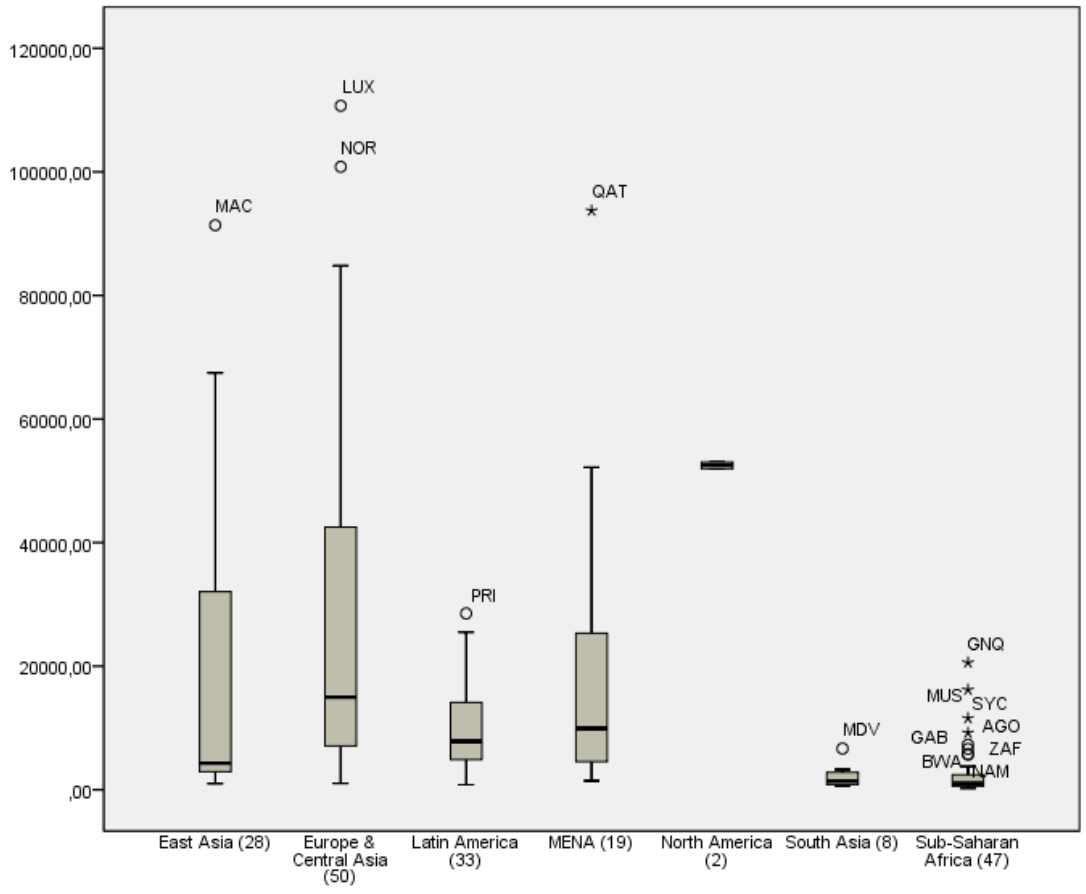
Temple (1998) argues that bad government policy outcomes are strongly associated with the low level of social capital, which is assumed to be shaped gradually throughout decades or even centuries. One of the most important dimensions of social capital is trust. The most influential recent study on the historical origins of mistrust in Africa is provided by Nunn and Wantchekon (2011). They find that individuals whose ancestors were more exposed to slave trade between 1400 and 1900 exhibit less trust toward unknown people and the government, but even toward relatives and friends. The intensity of slave trade could degrade trust through several channels. First, since

[2] Since in the empirical research the singularity of Africa appears in the form of a negative dummy variable which significance is hard to eliminate, some refer to this research direction as the 'quest for the Africa dummy' (Jerven 2011a, Englebert 2000).

the grow of demand in slaves increased the risk to be given away by friends and relatives and not only by raiders, untrusting behavior became a more rewarding rule of thumb in social interactions than trusting. Moreover, this practice is very likely to have been taught from generation to generation even after the slave trade was abolished. Second, slave trade might have harmed legal institutions which were supposed to deter from destructive social behavior.

The other branch of the literature explains harmful policies and policy outcomes with the low level of state legitimacy. The colonial states exogenously introduced in the 1880s were alien to the traditional African political structures and practices. Englebert (2000) argues that the westernized elite, which aimed to keep their privileged political position after independence, were encouraged to provide less public goods and increase trade barriers to provide the necessary resource for maintaining their patron-client system.

Figure 1.1 The boxplot of GDP per capita per regions (current USD, World Bank, 2013)



Note: Regions are defined according to the World Bank's classification scheme. The number of countries per region is shown in parentheses on the horizontal axis. Outliers and extreme outliers are labeled with circles and stars, respectively. Outliers are identified with country codes.

The nature and the duration of European colonization and the nationality of the colonizer are also found to be important sources of development differences across the

world. Former settler colonies where the number of Europeans was relatively high tend to perform better today in various aspects. La Porta et al. (1999) find that common law countries (British legal origin) have better quality governments than those with civil (French legal origins) and socialist law. Moreover, Olsson (2009) finds positive relationship between the duration of colonization and the level of democracy. Countries where Europeans remained the minority during the time of colonization experience higher income inequality (Angeles 2007) and more corruption (Angeles and Neanidis 2014) today. One of the most influential works on the development impacts of colonization is Acemoglu et al. (AJR, 2001), which investigates the effects of institutions on economic performance. They argue that where they could not settle due to unfavorable disease environment, Europeans set up extractive institutions which are associated with bad current institutions and poor economic performance. In their paper, the disease environment is measured with the mortality rates expected by the first European settlers proxied with the mortality rates of soldiers, bishops and sailors stationed in the colonies between the 17th and 19th centuries. The quality of the data and the final conclusion of this seminal work have been criticized by various scholars. Albouy (2012) concerns the reliability and comparability of the underlying dataset. Auer (2013) argues that AJR overestimates the importance of institutions in explaining economic growth and shows that the disease environment has a direct effect on prosperity. Bolt and Bezemer (2009) show that the argument established in AJR is not observable on a sample of African countries and the disease environment seems to effect long-term African economic performance through colonial education rather than colonial extractive institutions.

Another strand of the literature, which is very much related to the branch discussed above, attempts to understand how strongly European colonization have affected local conditions and the fate of indigenous societies. Were local cultures and conditions destroyed by these external impacts? Or, have indigenous societal features survived? Has European influence caused persistent transformation in indigenous socio-economic organizations or did it affect only in the short run? Based on a global sample of former European colonies, the 'reversal of fortune' hypothesis (Acemoglu et al. 2002) argues that territories which were relatively rich around 1500 are now relatively poor. Their explanation is the following: while Europeans were more likely to settle down on sparsely populated areas where they established good (property right-friendly) institutions, they did not settle on densely populated areas but exploited them. This thesis has been challenged from various aspects, for instance, for the oversimplification of the feautures of certain regions, for instance Africa (Austin 2008).

In relation to Africa, where the colonial period did not last long, recent studies have pointed out that precolonial norms and culture are still in operation and explain differences in contemporary institutions and development. Using anthropological data, Gennaioli and Rainer (2007) find that the quality of public goods is positively associated with the share of precolonial centralized groups within the current

population. They argue that the British relied on the traditional political arrangements in implementing colonial programs. Hence, in societies which were centralized before the colonial era, regional chiefs appointed by the British were accountable to the higher traditional authorities. Since, due to the lack of such institutions, the appointed chiefs of fragmented and non-centralized groups were directly subjected to the distant colonial administration, they often exploited their subordinates. This explanation is strongly linked to the argument of Englebert (2000) on the legitimacy of African states discussed above. Differences in the level of precolonial centralization explain economic performance discrepancies within countries as well (Michalopoulos and Papaioannou 2013, Bandyopadhyay and Green 2012).

Although some of Africa's mysteries have been solved, economics and development studies have discovered additional puzzling phenomena related to the continent. Some factors and historical events have influenced African societies in different ways than the other parts of the world. Nunn and Puga (2012) show that while in general irregular terrain affects economic possibilities negatively through making agricultural cultivation and transportation difficult, ruggedness in Africa had a beneficial consequence in the past: it protected certain groups from being raided for slave trade. Dincecco et al. (2014) find that the long term consequences of historical conflicts are unique in Sub-Saharan Africa. While the frequency of conflicts in the past predicts higher current state capacity everywhere in the Old World (their sample does not contain the Americas) including Africa, it is only Africa where past and modern time conflicts are significantly related.

Finally, we outline some current issues in African development research. As the above discussion illustrates, economics has recognized the role of history in determining development trajectories of African societies. However, there is a considerable debate on the method that economics uses to investigate development issues and the reconciliation of the the historical and economic approach. While historical analysis is usually based on case studies and qualitative evidence, economists attempt to arrive at generalizable results based on quantitative information and statistical methods. In his article, Hopkins (2009), while welcoming the freshness and boldness that economics (for instance Acemoglu et al. (2001) has brought into the research of African economic history, he criticizes the oversimplification (treating hundreds of years without taking the possibility of change into account, not recognizing important differences between colonies which have been identified by historians) that it sometimes makes due to unavailable or unsufficient data.[3] Improving the quality of historical data and collecting new information for empirical research in relation to developed as well as developing regions of the world, including Africa, have been one of the main concerns of economic historians in the past

[3] For more on this debate, consult Jerven (2011b), Fenske (2010, 2015), and Crafts (2012).

decade (Bolt and Green 2015, Greyling and Verhoef 2015, Frankema and van Waijenburg 2012, Frankema 2010). These new data provides new insight and sometimes surprising results on African economic history and development issues.

1.2.2 The linguistic features of Sub-Saharan Africa

Having discussed the peculiarities of Sub-Saharan Africa in various aspects, it should not be surprising that this region is also characterized by a unique language situation. This sub-section compares the linguistic patterns of Sub-Saharan Africa to other areas of the world and discusses how they are related to the historical and development issues reviewed above.

According to the 17th edition of Ethnologue (Lewis et al. 2014), Africa (North Africa and Sub-Saharan Africa is not reported separately), where we find about 15% of the world population, hosts more than the 30% of the world's languages (Table 1.1). While Asia gives about 60% of the world population, its share from the number of languages is only slightly more than Africa's. Regarding the average number of speakers, Africa is the third out of the five regions after Europe and Asia. Figure 1.2 displays the distribution of languages per status in each world region (Lewis et al. 2014). Northern and Sub-Saharan African countries are presented separately. 'Institutional' languages are recognized officially at the national or regional level or used or taught in the formal education system. 'Developing' and 'vigorous' languages are not recognized officially but have a sustainable position since they are actively used by all generations. The only difference between the two is that 'developing' languages are standardized. Languages 'in trouble' are losing speakers and/or spoken by older generations only. Languages which have rather symbolic value but only barely used for communication belong to the 'dying' category. (More description on the language status measure is provided in Chapter 2 and Appendix 2A).

Table 1.1 The distribution of world languages by area of origin

Area	Living languages		Number of speakers			
	Count	Percent	Total	Percent	Mean	Median
Africa	2,138	30.1	997,320,660	14.7	381,316	27,500
Americas	1,064	15.0	946,060,483	13.91	48,428	1,160
Asia	2,301	32.4	4,086,262,000	60.1	1,642,605	12,000
Europe	286	4.0	735,669,330	10.8	5,727,252	35,600
Pacific	1,313	18.5	35,284,389	0.5	5,166	950
Total	7,102	100.0	6,800,596,862	100.0	885,834	7,000

Source: http://www.ethnologue.com/statistics [26 Feb 2015]

Each region is characterized by a unique pattern. The share of 'vigorous' languages is the highest and the share of 'in trouble' and 'dying' category is the lowest in Africa. This observed pattern suggests that the majority of languages without official recognition and standardization are not likely to disappear in the near future. In

contrast, more than one third of the languages in Asia, Europe and the Pacific region are expected to die out within a few generations. The level of endangerment is the highest in the Americas: more than 60% of currently living languages are 'in trouble' or 'dying'. The language pattern of North Africa with the high share of languages at risk is more similar to that in the Americas.

We argue that the language patterns outlined above are strongly related to the historical experience with centralized states and the duration and nature of European colonization discussed in Section 1.2.1. Currently, Europe consists of nation states which came into existence as a result of a long-lasting historical process. Since state centralization and nation-building are usually accompanied by a preference for a single language which represents unity and contributes to the efficiency of the economy and government administration, the great majority of European countries acknowledge only one official language. If states oppress certain languages or the socio-economic advantages of the single official language are high enough, people are likely to abandon their primary language. In nation states, ethnic and linguistic borders are highly corresponding with country borders and the minority groups are usually bilingual in their mother tongue and the official language.

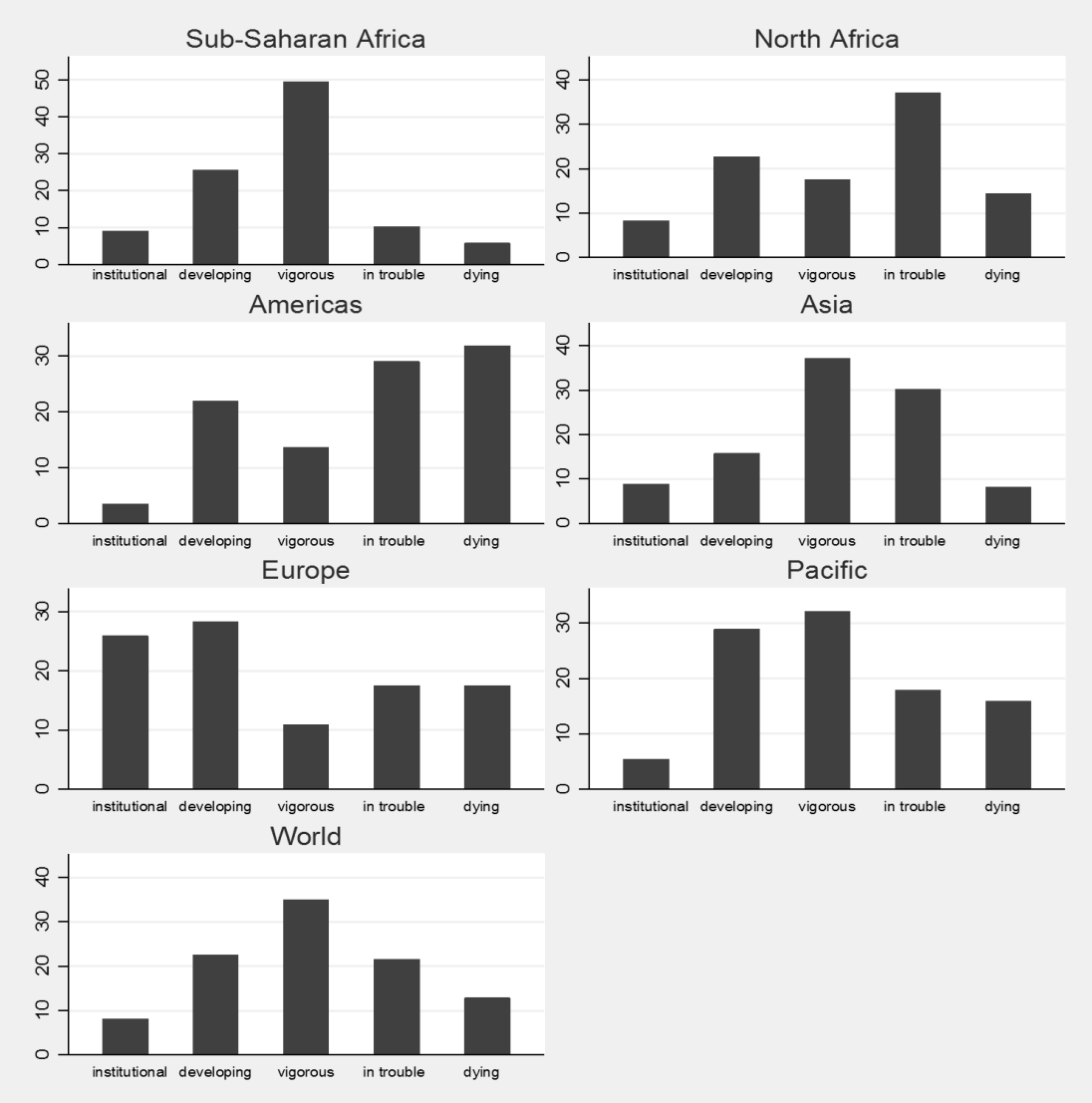
The share of endangered languages in Asia, as a whole, is similar to that in Europe but regionally more heterogeneous. In Eastern and South-Eastern Asia, which are historically characterized by relatively stable kingdoms with centralized army, religion and written languages, the majority of people speak the official language as either primary or second. In South Asia, which consisted of small kingdoms before the British conquest, the corresponding proportion is much lower and the distribution of languages across status categories is more similar to that in Sub-Saharan Africa.

In the Americas, where the colonial era lasted for centuries, the language of the former colonizers' are used as almost exclusive primary language by all ethnicities or races, therefore, as highlighted by Figure 1.2, native languages are highly endangered.

By the 8th century, North Africa was under the control of the Arab caliphate. Due to its prestige as the language of trade and Islam, Arabic has gradually crowded out the majority of local languages and now, except for Morocco, its standard form is the single official language in the African countries by the Mediterranean Sea. By the time of the Berlin Conference (1884-85), Sub-Saharan Africa hosted a huge number of societies that varied in size and the level of centralization. According to historical, anthropological and biological evidence, these societies were in active interaction through trade, migration, marriages and warfare (discussed in Chapter 2). Centralized states similar to those in Eastern and South-Eastern Asia were not present, and the local tribal kingdoms had limited size and power. Moreover, ethnicity and nation, which had a central place in European thinking by the 19th century, were unknown and meaningless terms there (van den Bersselaar 1997, Harris 1988). As for the colonial past, Africa hosted much less European settlers than North America and Australia, and for a much shorter period than Latin-America (Olsson 2009). Therefore, forces which have led to the extinction of smaller languages and the spread of a single dominant

language (either indigenous as in Asia and Europe or foreign as in the Americas) in other parts of the world were less intense or less effective in Africa. Although the languages of the former colonizers were used in public affairs and education during the colonial era and introduced as official in all countries after independence, proficiency in English, French, Portuguese, Italian and Spanish is considered relatively low (Chapter 3 provides more information on the distribution of European languages in Africa). Despite the lack of official recognition, precolonial trading languages still have high prestige and important role as lingua francas. The only comparable region is the Pacific where English and French are official but only a relatively small fraction masters them.

Figure 1.2 The distribution of language statuses per world region (in %)



Source: Ethnologue (Lewis et al. 2014, online version)

1.3 Language and society – a general overview

The aim of this section is to provide an overview on our current knowledge of the relationship between languages and the society. There are two options to accomplish this task. The first one is to list all related disciplines (mainstream economics and language economics, political science, anthropology, ecology, and sociolinguistics) and discuss their approach and main findings. However, since certain topics appear in various fields, redundancy in this type of discussion would be unavoidable. Instead, we choose the other option and provide an overview on the topics that are closely related to the issues covered in the chapters of the current thesis.

1.3.1 The socio-economic impacts of ethnolinguistic diversity

The roots, socio-economic consequences and measurement problems of ethnolinguistic diversity, which is also referred to as heterogeneity, fragmentation, or fractionalization in the literature, has been the concern of many scholars of different fields in the past decades.

Economics views ethnolinguistic diversity as an important cause of underdevelopment and poor quality institutions. The earliest studies are related to the research aiming to solve the mystery of Africa's underdevelopment (discussed in Section 1.2.1). Mauro (1995) estimates the causal effect of corruption on economic growth on a cross-section of countries and uses ethnolinguistic fragmentation (diversity measures are discussed in Section 1.3.2) as an instrument to handle the potential endogeneity between corruption and growth. He finds that ethnolinguistic fragmentation is positively associated with corruption, which lowers investments and, therefore, undermines economic growth. This was the first empirical work to reveal an indirect channel through which diversity affects macroeconomic performance. Investigating the causes of Africa's growth tragedy, Easterly and Levine (1997) argue that while the indirect influence of ethnolinguistic diversity on growth through its harmful effect on public policies (insufficient provision of education and infrastructure) is quite robust, the direct effect is more ambiguous. Following these early works, ethnolinguistic diversity has gradually become a popular variable in empirical models investigating cross-country differences in economic growth (Collier 2000), socio-economic development (Gerring et al. 2015, Tequame 2010, Putnam 2007), public good provision (Alesina et al. 1999), institutional quality (Aghion et al. 2004, La Porta et al. 1999), and internal conflict (Montalvo and Reynal-Querol 2005). However, some emphasize (Alesina et al. 2003) that the empirical results on the relation between ethnolinguistic diversity and the aspects of development are sensitive to the properties of the sample, the methodology and the choice of variables. There is a considerable number of articles which find that the effect of heterogeneity on socio-economic outcomes is insignificant (Gerring et al. 2015), not convincing (Fish

and Brooks 2004), even positive (Arcand and Grin 2013) or dependent on other factors (Easterly 2001, Collier 2000).

Although the previous paragraph highlights the contribution of economics to the ethnolinguistic diversity literature, at this point it is necessary to note that sociolinguistics recognized the above explained relationship decades before economics. Greenberg (1956) suggests a standardized set of diversity measures to enable a relatively objective comparison between regions which might serve as an appropriate basis for empirical research seeking the relationship between linguistic diversity and political, economic, geographic, and historic factors. Pool (1972) correlates GDP with linguistic heterogeneity and concludes that while linguistically homogeneous countries could be poor, heterogeneous ones could never be rich.

Having recognized the societal importance of heterogeneity, the question naturally rises: why are some countries more fractionalized than others? Anthropological and economic literature provide convincing evidence that the territorial distribution of ethnolinguistic groups is associated with geographical and climatic conditions such as proximity to the equator and variability in land quality, elevation and climate (Kaufman 2014, Green 2013, Michalopoulos 2012, Cashdan 2001, Nettle 1999). There are two popular theories to explain this relationship. Nettle (1999) suggests that where conditions allow for a longer growing season, self-sustaining is relatively easy and cooperation among groups is not necessary. Higher risks in food supply encourage interaction between groups which is expected to result in the formation of a common language in the long run. While Nettle's theory is based on the necessity of cooperation among multiple groups, the starting point of the other theory introduced by Ahlerup and Olsson (2012) is a single, undivided group. They argue that if the provision of collective goods is insufficient, populations located at the peripheries are likely to detach and form new groups which lead to a genetic and cultural (linguistic) drift over time.

Ethnolinguistic diversity is present everywhere in the world, however, in a different form. According to its origins, there are three types of ethnolinguistic diversity. The first one, which is extensively discussed above, is related to traditional societies where different cultures have been in a dynamic relationship for centuries. Second, due to the increasing immigration from developing countries in our globalized world, developed countries face the challenges of cultural and linguistic heterogeneity. And finally, the establishment and enlargement of international communities, such as the European Union, has to cope with the problem that the citizens of the member countries speak different languages and exhibit different language learning behavior. This type of heterogeneity can undermine the success of policies that aim to create a single, integrated market. Since the latter two types are not closely related to the topic of the thesis, they are not discussed in more detail.

1.3.2 Diversity measurement[4]

The main goal of the previous section is to list the areas where the effects of ethnolinguistic diversity have been researched. But before we start to discuss the types of indicators applied in various fields, it is necessary to devote a few sentences to the discussion of the term 'ethnolinguistic'. The conceptualization and the measurement of this concept are quite problematic.

First, the term 'ethnolinguistic', which is used in the earliest studies (Mauro 1995, Easterly and Levine 1997) and often in the latest ones (Wang and Steiner 2015) as well, implicitly assumes that ethnicity and language refer to the same concept or ethnic and linguistic heterogeneity have the same societal consequances, thus distinction between the two is not necessary. Although ethnicity and language are related concepts (Chandra 2006) and in practice these terms are used interchangeably (Laitin 2000), the linguistic and ethnic composition of a society might differ significantly (Chapter 3 of this dissertation, Anderson and Paskeviciute 2006). However, when the results of the existing literature are discussed (for instance in the Introduction) and in Chapter 2, in which available data do not allow for the distinction between ethnicity and language, this dissertation also applies the term 'ethnolinguistic'. Since they are based on survey data which contain information on language and ethnicity separately, Chapter 3 to Chapter 5 make a distinction between 'ethnic' and 'linguistic' diversity.

The second issue related to the conceptualization and measurement of diversity is that scholars have soon recognized that one approach does not fit all purposes. The following overview demonstrates the evolution of diversity measurement.

The simplest possibility to capture the heterogeneity of the society according to one characteristic (e.g. language, ethnicity, religion) can be defined as shown in Eq. 1.1.

$$A = 1 - \sum_{i=1}^{N} p_i^2 \qquad \text{(Eq. 1.1)}$$

where i is the indication of the group, p_i is the share of group i within the society, N is the number of the groups, and $\sum_{i=1}^{N} p_i^2$ is the Herfindahl-index (Herfindahl 1950) with $\sum_{i=1}^{N} p_i = 1$. Measure A refers to the probability that two members of the population selected at random belong to different (linguistic, ethnic or religious) groups. The Ethnolinguistic Fragmentation (ELF) (Taylor and Hudson 1972), the first measure used in early empirical works such as Mauro (1995) and Easterly and Levine (1997), is based on this concept. However, it became a subject of criticism in the 2000s for a number of reasons. First, as some highlight (Alesina et al. 2003), although

[4] More detailed overview on existing measures are found in Gisselquist and McDoom (2015) and Ginsburgh and Weber (forthcoming).

it is labeled 'ethnolinguistic', it utilizes mostly linguistic data drawn from the Atlas Narodov Mira compiled by Soviet scholars in 1964 (Bruk and Apachenko 1964). The improvements achieved in the following years include the utilization of more up-to-date data[5], the distinguishing between ethnicity and language (Alesina et al. 2003, Fearon 2003), and the broadening of the range of countries for which the measure is available (from about 130 in Taylor and Hudson 1972 to about 200 in Alesina et al. 2003 and Fearon 2003). Second, by elaborating the PREG (Politically Relevant Ethnic Groups) index, Posner (2004a) draws the attention that different research questions require different approaches. He shows that a diversity measure that focuses only on those groups which play a significant role in policy making are more applicable for studies that highlight the impact of ethnolinguistic fragmentation on economic development through shaping political actions. And third, recent studies criticize the above mentioned measures for treating diversity as a static characteristic of a society. Regularly repeated surveys (national censuses, World Value Surveys) containing individual level data are suggested to control for dynamics in diversity (Koster 2012, Fedderke et al. 2008, Campos and Kuzeyev 2007).

The second way to measure heterogeneity (Eq. 1.2) is also based on grouping the members of the society along one possible dimension, but it recognizes that some groups are more similar than others, and more distant groups contribute to the diversity of the society to a larger extent.

$$B = 1 - \sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j r_{ij} \qquad \text{(Eq. 1.2)}$$

where B is the diversity measure, i and j are indications for the groups, p_i and p_j are population shares of group i and j within the society, N is the number of the groups, and r_{ij} refers to the elements of a similarity matrix showing the resemblance between group i and j. The difficulty of this measure lies in finding the appropriate tool to capture group distances. The two most popular tools for proxying linguistic distance are the glottochronology-based method (Greenberg 1956, Dyen et al. 1992, Desmet et al. 2009) that gives information on the share of basic vocabulary in common and the tree diagram (Grimes and Grimes 1996, Fearon 2003, Laitin 2000) that represents the structural relation between languages.

However, it is not only the linguistic distance between groups that might be relevant for economic concern. Baldwin and Huber (2010) introduce the Between Group Inequality (BGI) index which takes the income distance between ethnic groups into account and show that it does a better job in explaining public good provision than

[5] World Christian Encyclopedia (Barrett et al. 2001 and 1982), Encyclopedia Britannica (2000), Ethnologue (Grimes and Grimes 1996), the CIA World's Factbook

Eq. 1.1 and the linguistic distance-adjusted in Eq. 1.2. Thus, Eq. 1.2 is able to consider how the distribution of a secondary feature might reinforce or weaken the impact of diversity measured along the primary dimension on economic and political outcomes of main interest. Or in other words, measure B is able to account for similarities between groups along a special dimension while still capture dissimilarities in terms of another.

It can be shown that on the country level, Measure A is a special case of Measure B when a continuous distance measure between groups is not possible, thus the elements of the similarity matrix might assume only 0 or 1.

Some recently developed measures approach ethnolinguistic diversity uniquely. Political and conflict studies favor the concept of polarization (first introduced by Esteban and Ray 1994) over fragmentation. A simplified version of the polarization index developed by Reynal-Querol (2001) is designed to show how far the distribution of groups (ethnic, linguistic, religious etc.) in a population differs from the bimodal distribution. The success of this index in explaining the probability and length of ethnic conflicts (Montalvo and Reynal-Querol 2010, Montalvo and Reynal-Querol 2005) suggests that it is not the structure of the society but rather the number and the size of competing groups that is relevant for the occurrence of civil conflicts. De Groot (2011) develops the Ethnolinguistic Affinity (ELA) that measures the similarity of the ethnic composition of neighboring countries to explain conflict spillover. Alesina and Zhukovskaya (2011) suggest that the geographical distribution of groups also should be considered when it comes to development issues. The segregation or isolation is expected to determine their possibility to represent their interest and the necessity to cooperate with other groups.

Some scholars argue that when it comes to heterogeneity measurement, one should account for all possible relevant dimensions of diversity that might be interesting for scientific consideration. The Social Diversity Index (Okediji 2011 and 2005), the Generalized Index of Fractionalization (Bossert et al. 2011), and the Distance-Adjusted Ethnolinguistic Fractionalization (Kolo 2012) attempt to fulfill this requirement in their own way.

And finally, there are some indicators that are designed for addressing linguistic issues in the expanding European Union. The communication potential[6] or Q-value elaborated by De Swaan (1993) provides a framework to determine which languages and language repertoires ensure communication with the highest proportion of EU citizens and show how the optimal language repertoire changed due to the admission of new member states between the 1960s and 1990s. The expansion of the EU in 2004 with ten new countries, which national languages are not widely spoken outside their borders and where the proficiency in foreign languages is relatively low, has raised the

[6] The Index of Communication Potential (ICP) elaborated in this thesis is different from the communication potential or Q-value by De Swaan (1993).

following dilemma. On the one hand, the European Union aims to secure the equality of citizens by recognizing their languages as official working languages of the community. On the other hand, translation and providing documents in 24 languages is expensive. A measure by Ginsburgh and Weber (2005) and Ginsburgh et al. (2005) calculates the proportion of disenfranchised people in the EU under different official language scenarios.

The aim of this subsection is to demonstrate the existing approaches to diversity measurement and help position the Index of Communication Potential (ICP), one of the main outcomes and tools of this thesis, in the system of the these approaches. As it is discussed in Section 1.4 and Chapter 3 to 5, the ICP is distinct from existing diversity indicators in certain aspects, while similar to those in other terms. On the one hand, the novelty of the ICP is that it is able to handle multilingualism and shows the probability that two people can communicate because they speak at least one common language. On the other hand, Appendix 5B shows that the formula of the ICP is a special case of Measure B, where p_i and p_j are the individual weights in the underlying sample and the distance matrix, r_{ij}, is assigned 1 if individual i and j have at least one language in common and 0 otherwise.

1.3.3 Language as capital and capability

Languages can facilitate or prevent that its speakers achieve certain goals. One of the main concerns of language economics is the role and the value of languages on the labor market, which is closely related to human capital theory, economics of education, labor economics, and well-being. Languages can be considered as a form of human capital (Grenier 1982) which value should be mirrored in wages; and also a marker of sociocultural identity which can serve as a basis for labor market discrimination (Pendakur and Pendakur 2002). The earliest studies concentrated on the competition between English and French in bilingual Canada (Carliner 1981, Vaillancourt 1980) and the language acquisition behavior of Spanish speaking immigrants in the US (Grenier 1984). Current empirical research on the relationship between language proficiency and personal income is available for a broad range of both developed (see Dustmann 1994 on Germany, Rendon 2007 on Spain, for instance) and developing countries (see Levinsohn (2007) on South Africa and Aldashev and Danzer (2014) on Kazakhstan, for instance).

Sociolinguists have revealed a wide range of situations in which minority language speakers are disadvantaged (Batibo 2005: 55-57). Children who are not proficient in the official and national languages in which schooling is provided usually face two types of difficulties. First, students who were brought up in a particular sociocultural environment find it difficult to adjust to the new social environment and are likely to drop out of the schooling system. Second, learning in an inadequately mastered language is accompanied by lower confidence and poorer performance in school.

Minority language speakers are also likely to be less able to participate in political decision making and in transactions (trade, bank loans) which require the knowledge of the official or dominant language. Since languages make it possible to disseminate information, minority language speakers are likely to know less about HIV/AIDS prevention, other epidemics, or modern farming methods.

Recent studies point out, that the enlargement of the European Union have raised some linguistic issues that are similar to those in the traditionally multilingual societies. Citizens of the youngest member states cannot exploit the possibilities offered by the single labor market due to the low level of foreign language proficiency (Aparicio Fenoll and Kuehn, forthcoming).

1.3.4 Language dynamics

The study of language dynamics is closely related to the study of language as capital or capacity. All scientific fields and models that aim to explain language dynamics accept that if the reward from speaking a language is not high enough (relative to another language), the language is expected to get abandoned and die out within a certain period of time (Lewis and Simons 2010, Batibo 2005, Fishman 1991). Thus, the unanswered questions are: what determines the value of languages and how the process of extinction is proceded.

Sociolinguistics suggests that the pressure that languages have to face can be caused by demographic superiority (higher number of speakers), socio-economic attractions (socio-economic opportunities such as higher wages), political dominance (language is associated with power and political influence), and cultural value (religious languages) associated with other languages. If weaker languages come into contact with stronger ones, its speakers want to identify themselves with the more prestigious or more valuable language. Moreover, there are certain circumstances (geographical distribution of groups, migration, interethnic marriages, political decisions, and the functionality of languages) that influence the spread of languages and the chance for contacting a stronger language (Batibo 2005: 93-94).

Language policy is one of the factors that can change the value attached to languages through influencing choices at the individual level and language dynamics at the society level. The possibilities of economic analysis in supporting language policy and planning have been one of the main concerns of language economics from the beginning (Zhang and Grenier 2013, Grin 2003). First, economics is eligible to analyze how the promotion or suppression of languages determines economic outcomes and vice versa (Muravyev and Talavera 2015). Second, the cost and benefit analysis can be applied for comparing language policy alternatives (Ginsburgh et al. 2005, Patrinos and Veles 1995, Vaillancourt 1995). And finally, certain economic concepts and theories such as microeconomic optimization theory, game theory and public choice theory might help decide what to do with languages and how in order to achieve certain social goals (Wickström 2014).

In order to reverse language shift and prevent the pauperization of minority language speakers, a number of academic and international language revitalization programs have been initiated throughout the world including Africa.[7] Language documentation, standardization and literacy programs are proposed as the main tools (Grenoble and Whaley 2006).

Some recent works, published in high-ranked science journals such as Nature and Physica A, attempt to model and simulate the competition of languages in a hypothetical world under simplified circumstances. The basic model (Abrams and Strogatz 2003) argues that in a bilingual society where the attractiveness of languages is dependent on their status and the number of speakers only one of the competing languages is expected to die out. Later studies (Patriarca and Leppänen 2004, Stauffer and Schulze 2005, Pinasco and Romanelli 2006, Patriarca and Heinsalu 2009) have shown that the outcome of such a competition is dependent on other factors (the number of competing languages, geographical distribution of speakers, population density, historical circumstances etc.) and extinction is not necessary.

1.4 Research goals, data, methodology and the outline of the thesis

1.4.1 Research goals and questions

Although the above presented review demonstrates that the scientific literature has revealed many aspects of the relationship between languages and the society, some gaps still remain, especially in relation to multilingual developing societies.

As it was highlighted earlier, the most researched area of the language situation in relation to development issues is ethnolinguistic diversity which is understood as the probability that two randomly selected citizens of the society speak different primary languages. Although some indicators handle the distance among languages, none of them account for other than primary languages.[8] Since the majority of the population in linguistically fragmented countries is multilingual, ignoring second languages, which are assumed to overcome linguistic barriers, could lead to somewhat biased results when analyzing the relationship between the language situation and socioeconomic outcomes (Laitin 2000). Although the necessity of diversity measures that can be applied for multilingual societies has early enough been acknowledged (Greenberg 1956), due to data availability problems, actual measures are difficult to construct. Aggregated information from Ethnologue (Lewis et al. 2014), for instance, is not exploitable for this purpose. Such a desired measure requires information on individual language repertoires or, at least, the share of each language repertoire

[7] For instance, the Hans Rausing Endangered Languages Project at SOAS (http://www.hrelp.org/) and the UNESCO's Endangered Languages Programme (http://www.unesco.org/new/en/culture/themes/endangered-languages/).

[8] Issues related to the definition of linguistic terms such as home language, mother tongue, first and second language and additional languages are discussed in Section 3.3.3.

within the society. One of the main goals of the thesis is to create a linguistic measure which accounts for the complete linguistic repertoire of individuals and can accompany or substitute existing ethnolinguistic fragmentation indices in development research (Goal 1). We utilize the unique dataset of the Afrobarometer Project (more details in Section 1.4.2 and Section 3.3) to accomplish this objective. The construction of such a measure implies the following questions. How to interpret a linguistic indicator that assumes multilingual individuals instead of monolinguals (Question 1)? What are the advantages of such an indicator compared to existing measures (Question 2)?

The most important functions of languages are communication, identity construction, and the transmission of culture and traditions. If we consider the communication function of languages in its strict sense (i. e. we ignore that languages are often attached cultural and symbolic value), it should not matter if information is transmitted in one's primary language or in another: in theory, they can be equally efficient (especially in simple situations). However, when it comes to the other two functions, we expect that the cultural and symbolic value attached to primary and secondary languages are different. While the social and political science literature provide evidence that the identification role and cultural value of primary languages is strong, the importance of additional languages in these aspects is less established. Thus, this thesis aims to contribute to our understanding of the societal functions (other than communicational) of multilingualism and second languages in Sub-Saharan Africa (Goal 2). This attempt can be broken down into two questions. First, do second languages have identification role (Question 3)? And second, can other than primary languages counterbalance the acknowledged harmful effects of ethnolinguistic diversity (Question 4)?

And finally, this thesis also attempts to relate to the research direction that seeks to reveal the roots of the current language situation (Goal 3). While existing studies predominantly focus on the geographical and climatic determinants of ethnolinguistic diversity across the globe (Section 1.3.1), we attempt to explain how historical factors have influenced a less investigated dimension of the language situation, namely language status (more on the concept is found in Section 2.3.1), in the Sub-Saharan African context (Question 5). Although it might not be straightforward at first, our results contribute to our understanding on the dynamics of ethnolinguistic diversity as well. As we highlighted above, the economic and social value of languages, which are captured in our language status concept, determine whether a group maintains or abandons a language. Consequently, since the (relative) size and the number of groups within a territory depend on these decisions, the factors that influence the value of languages are assumed to play an important role in shaping the dynamics of ethnolinguistic diversity.

1.4.2 Data

The language-related data in this thesis are obtained from two sources. Chapter 2 is concerned with the long-term determinants of the current status of languages which is measured with the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Lewis and Simons (2010)) from the 17th edition of Ethnologue (Lewis et al. 2014). EGIDS is originally designed to evaluate the level of language endangerment and to identify the most problematic areas where actions are required in order to reverse language decline. Based on their official status, established orthography, communication role, intergenerational transmission, and identification (symbolic) function, languages are grouped into 13 classes where higher values label more disruption, i. e. higher probability for language extinction in the short run. More description is provided in Section 2.3.1 and in Table 2A.1. EGIDS has two main advantages as a measure for language status. First, as it is highlighted above, it is a multidimensional measure which takes several aspects of the status of languages (endangerment) into account. And second, it is more refined than any other existing classification schemes (the Graded Intergenerational Disruption Scale (GIDS, Fishman 1991), the UNESCO's language endangerment classification scheme (UNESCO 2003, Moseley 2010), and the language status measures applied in previous editions of the Ethnologue).

Chapters 3 to 5 use the language- and ethnicity-related variables of the 4th Round the Afrobarometer Survey.[9] The Afrobarometer is an independent, non-political research survey on the social, political, and economic status of citizens in a growing number of Sub-Saharan African countries[10]. While the Afrobarometer has become a popular source of data in development research (see for example Eifert et al. 2010, Nunn 2010) it is only Round 4, conducted in 2008 and 2009 in twenty countries[11], that collects data on home and other spoken languages. Round 4 contains 27713 observations in total with 1200 as typical sample size per country. The three largest countries in the sample, South Africa, Nigeria and Uganda, cover 2400, 2324, and 2431 citizens respectively. A more detailed description on the Afrobarometer Survey is found in Section 3.3, Section 4.3.1, and Section 5.3.

The key linguistic variables from the Afrobarometer are Q3 on home languages, Q79 on ethnicity and Q88E on second languages. While respondents were required to select their ethnicity and home language from a predefined list, additional languages in Q88E are self-reported. One of the drawbacks of self-reported language proficiency is that the same language may be known under different names, thus there is a risk of double

[9] http://www.afrobarometer.org [27 March 2015]
[10] 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 that covers 36 countries including those in Northern Africa is being processed and digitalised at the moment. Round 6 is under preparation.
[11] Benin, Botswana, Burkina Faso, Cape Verde, Ghana, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mozambique, Namibia, Nigeria, Senegal, South Africa, Tanzania, Uganda, Zambia, Zimbabwe

counting. The 16th edition of Ethnologue (Lewis 2009) was applied to identify names that refer to the same language. The Ethnologue is a comprehensive catalogue which provides information on the number of speakers, the geographical location, official status, endangerment and general references on world's languages. The EGIDS which is discussed above is published in the 17th edition. The latest, the 18th edition, which contains some changes compared to previous editions, was launched on 21st February 2015, after the finalization of the chapters of this thesis. These language-related variables serve as the basis for the Index of Communication Potential. The benefits and limitation to use Afrobarometer as a language data source is discussed in Section 3.3.

Demographic data on individuals in the Afrobarometer are also utilized extensively throughout the thesis. Q84C on trust in unknown people is the dependent variable of the empirical analysis in Chapter 4. Q83 on the importance of national versus ethnic identity plays a similar role in Chapter 5. Both of these chapters rely on additional demographic information such as age (Q1), gender (Q101), education (Q89), religion (Q90), living circumstances (Q8A-Q8E) and the location of residence (URBRUR).

Chapter 2 relies heavily on historical and geographical data. One of the main sources is the Ethnographic Atlas (Murdock 1967) which aims to provide information on the precolonial social and economic characteristics of 862 societies across the world including about 290 African tribes. The variables included in the atlas concern the characteristics of the economy, the type of agriculture and animal husbandry, family and community organization, gender roles, inheritance practices and the presence of different occupations. Murdock's Ethnographic Atlas serves as a base for two other databases that are, alongside the original, extensively used in the development literature. The first one is Gray's (1999) Corrected Ethnographic Atlas which contains more societies than Murdock's work and includes some corrections and additional variables.[12] The second one is the Standard Cross-Cultural Sample (SCCS, Murdock and White 1969) which contains only the 186 best described cultures from the Ethnographic Atlas across the world and about 2000 variables. The SCCS is continuously extended and published in World Cultures.[13] The variables included in the above mentioned sources are often applied in empirical studies explaining the global or regional distribution of development and certain societal characteristics (Rijpma and Carmichael 2015, Michalopoulos and Papaioannou 2013, Gennaioli and Rainer 2007). In Chapter 2, the socio-economic development of African societies is measured with five variables from Murdock's Ethnographic Atlas (the intensity of agriculture, jurisdictional hierarchy, class stratification, succession of the office of local headmen, and the presence of high god in the local religion) that are measured on an ordinal scale and are assumed to reflect the socio-economic complexity of a society

[12] Although Chapter 2 contains only African societies that are included in Murdock's Ethnographic Atlas, the final version of the paper is planned to be extended with groups covered in Gray's Corrected Ethnographic Atlas.

[13] http://eclectic.ss.uci.edu/~drwhite/worldcul/world.htm [21 July 2015]

(Section 2.3.3). Although the Ethnographic Atlas intends to provide information on societies when they were independent and not influenced by external cultural effects (e. g. European colonization), we show that it is very likely to fail on this effort (Section 2.4.2).

Chapter 2 utilizes several other databases. The Roome map (1924) shows the location of Christian missionaries in the early 1920s. The dates of the first Bible translations are derived from three sources (Groves 1964, Lewis et al. 2014, worldbibles.org). Geographical and climatic variables are from Fenske (2014), White (1983) and other online available sources listed in Section 2.3.5.

1.4.3 Methodology

While each chapter applies different methods to answer the underlying research questions, one common feature is that I apply statistical methods to arrive at fundamentally qualitative results.

In Chapter 2, the main methodology is the counterfactual analysis, based on regression analysis. The main aim of this method is to overcome the problem that historical experiments cannot be conducted, thus questions like 'what would have been if…?' cannot be answered. The counterfactual analysis has been made popular by Fogel (1964) who estimated the effect of railways on the economic growth in the 19th century USA by using a counterfactual situation of an USA without railways as the basis of comparison. Similarly, in Chapter 2, I attempt to reconstruct the current patterns of the status of languages if there were no European colonization and Christian missions in Sub-Saharan Africa. By comparing the reconstructed and the actual situation, I conclude that the early presence of Europeans in Sub-Saharan Africa still has a significant impact on the current language status and language situation in Africa. The languages of those ethnic groups which were preferred by Christian missions are likely to have a disproportionally higher degree of development than the rest.

Relying on the language- and ethnicity-related variables of the Afrobarometer Survey and the Index of Communication Potential (ICP), Chapter 3 provides a detailed description and comparison of the language situation (including ethnic and linguistic fragmentation) of the twenty countries included in the Afrobarometer Survey. We show how a simple graphical representation of the ICP can be employed to reveal the most important dimensions (e.g. the average number of spoken languages, the role of languages in communication potential, and the presence of monolingualism) of the language situation. This tool is helpful for language policy analysis on a continental scale, which so far has been made on country or regional level.

In the final two chapters, I apply multilevel or hierarchical modelling to find out if individual trust toward non-family members (Chapter 4) and national identity (Chapter 5) can be explained by the ability to communicate with others. The main benefit of the multilevel method is that it is able to take into account the effect of the

sociocultural environment on individual behavior within a single empirical framework. Or, in other words, the method can separate the effect of individual and environmental characteristics on the dependent variable. In statistical studies, it often happens that the impact of the same factor on the investigated phenomenon is essentially different if that factor is measured at the level of individuals and when it is averaged on, for instance, the regional or country level. Thus, the aggregation of information can influence the results of the statistical analysis. When this is not taken into account and the results based on aggregated (group or country-level) data are applied to explain individual behavior is often called the ecological fallacy (Robinson 1950). The hierarchical model is a method to avoid this trap. In both chapters, two-level models are estimated with individuals as level-1 units. However, while in Chapter 4 the administrative regions serve as level-2 units, ethnic groups play the same role in Chapter 5. An additional difference between the two chapters is that while empirical models in Chapter 4 contain level-2 variables (such as ethnic, linguistic, religious and political fragmentation, regional well-being, inequality, average corruption, the share of citizens participating in voluntary organizations, share of Christians, population density, average ICP), models in Chapter 5 are more simple and do not contain ethnic-group level variables. More description on the benefits and application of the multilevel modelling is provided in Section 4.3.1.

1.4.4 Results and the outline of the thesis

Based on the linguistic information in the Afrobarometer Survey, we develop the Index of Communication Potential (ICP) which, according to the author's knowledge, is the first linguistic indicator that captures the multilingual nature of societies and available for multiple countries.

Since it is based on individual level information, the ICP can be computed at any desired level of aggregation. Individual ICPs are understood as the probability that a certain person can communicate with a randomly selected other person from the society given their language repertoires. Individual-level indicators, then, can be aggregated (averaged) at higher (e.g. region, country, urban-rural distinction) levels. The country-level ICP, for instance, is the probability that two randomly selected citizens can communicate with each other since they have at least one common language. Technical details are found in Appendix 3A and Appendix 5B. Moreover, Appendix 5B also shows that the ICP is a special version of the Herfindahl-index (which is the basis for the majority of diversity indices). In its original form, the ICP focuses on the linguistic similarity of citizens. However, when we deduct the indicator from one (1-country-level ICP), the value can be interpreted as the probability that two randomly chosen people do not have any common languages. The reversed form of the ICP is more similar to the concept of the existing diversity indices which also highlight linguistic dissimilarities within societies.

Chapter 3 (forthcoming in African Studies (Taylor and Francis)) is a rather descriptive part of the dissertation relying highly on the Afrobarometer data and the Index of Communication Potential. The chapter aims to compare and reveal the most important aspects of the language situation in the 20 sample countries in a way that might be useful to a wider audience in social and political sciences. The benefits and limitations of the Afrobarometer as a linguistic data source and its comparison with other available materials are extensively discussed. We show that although the Index of Communication Potential tends to be higher in countries with lower ethnic and linguistic fragmentation, the relationship is not deterministic. In certain countries, high diversity is accompanied with high average ICP. This suggests that taking other than home languages into account indeed leads to different linguistic patterns compared to those based on existing diversity measures described in Eq. 1.1 and Eq 1.2 in Section 1.3.2. Investigating multilingualism alongside diversity could open up promising research directions in development studies. Moreover, we show how a simple graphic representation of the ICP can be utilized to evaluate and classify the language situation in the sample countries. We show which languages contribute to the ICP to the largest extent (which have the most important communication value), how many languages the representative citizen speaks, which groups are likely to remain monolingual, the spread of the former colonizers' languages, and the relationship between indigenous and officially recognized European languages. If underlying data are available, the graphical representation of the ICP is eligible to monitor if language dynamics correspond with language and education policy goals and to forecast changes in language use. The chapter is accompanied with a supplementary material which contains bibliography on the language situation and policy per country and juxtaposes available linguistic data from various sources.

Chapter 4 (published in Social Science Research (Elsevier), vol. 49, pp. 141-155.) and Chapter 5 (forthcoming in M. Gazzola and B-A. Wickström, eds., The economics of language policy, Cambridge: MIT University Press) investigate the effect of multilingualism measured with the ICP and the number of spoken languages on two extensively researched social phenomena, notably generalized trust and national identity, which are found to be affected negatively by high ethnolinguistic fragmentation. Empirical research provides evidence that diversity is associated with low social capital and trust both in developed countries facing increasing immigration (Gerritsen and Lubbers 2010, Hooghe et al. 2009) and developing countries (Mavridis 2015). Although individual ICPs do not seem to influence generalized trust in Chapter 4, people living in an administrative area with higher average ICP are likely to trust unknown people more. Moreover, we also find that linguistic and ethnic fragmentation is positively associated with generalized trust, which supports the theory that if different group have intense interaction, they get accustomed to diversity and accumulate positive attitudes towards each other in the long run (Stolle and Harell 2013). Chapter 5 reveals the positive effects of second languages on another aspect of social cohesion. The failure of nation-building policies in post-independence African

states is often assigned to high cultural and ethnolinguistic fragmentation (Bannon et al. 2004). While we do not find evidence that ICP is associated with national identification, people speaking more than two languages turn out to be more committed to the nation than their ethnic groups. Thus, Chapter 4 and Chapter 5 suggest that multilingualism actually can be beneficial for social cohesion in the linguistically heterogeneous Sub-Saharan African countries.

As discussed in Section 1.3.4, the causes of languages dynamics has been researched by various disciplines. Sociolinguistics argues that the ultimate cause of the decline of languages is the change in their socioeconomic value. People are likely to maintain and acquire languages that enable them to gain higher standards of living. Language policy attempts to alter the socio-economic value of languages to influence individual language choices. However, as Section 1.2 and Chapter 3 show, language dynamics are not always determined by formal language policies. Chapter 2 (published as CGEH Working Paper No. 66, Utrecht University) attempts to reveal how the special characteristics and historical events of Sub-Saharan Africa shaped its current language status patterns. Unlike the other chapters which predominantly rely on the Afrobarometer data, Chapter 2 covers 389 ethnolinguistic groups located in 47 countries and derives information from various historical and contemporary sources. The theoretical framework is based on the findings of the historical, ecological, anthropological, economic, and political literature. Our main results are the following. Pre-colonial conditions are found to be important determinants of the current language status. Local societies that were more centralized in the 19th century are more likely to have developed and officially recognized languages today. The role of missionary activities is crucial: earlier Bible translations indicate higher status today. The nationality of the former colonizer has only indirect effect through regulating missionary activities. The indirect effects of geography on the status of languages through determining the location of missionaries and the development of indigenous societies are found to be more important than its direct impacts. Early European contact measured with the intensity of early missionary activities contributed to higher inequality between local groups both in terms of language status and socio-economic development. Table 1.2 shows the outline of the thesis.

Although it is not a separate chapter, the nationality of the colonizer is a reappearing issue throughout the thesis (Chapter 2, 3, and 5). As the historical literature summarized in Chapter 2 suggests, the British, French, Portuguese, and Belgian colonizers treated indigenous languages differently. We find empirical evidence that the relationship between European and local languages, the spread of European languages and the association between the language situation and socioeconomic development is dependent on the nationality of the colonizer. However, as it is common in the literature, we predominantly focus on the British-French distinction. As Chapter 2 suggests, the nationality of the colonizer seems to have an indirect impact on the current status of languages through its effect on missionary activities. The competition between Christian dominations and the intensity of early

activities were higher on former British territories. Based on Chapter 3 and 5, we arrive at the following conclusions. First, the proficiency in the colonizer's language is lower in former French colonies. Second, if there is a local wide-spread language, people are more likely to acquire that instead of the former colonizer's language in former French colonies. English rather complements the dominant local language than substitutes it in former British colonies. And finally, while languages do not play a significant role in building national identity in ex-British colonies, higher ICP is associated with stronger national commitment in ex-French colonies.

Table 1.2 The outline of the thesis

	Thesis			
Main issue	Roots	Measurement	Development impacts	
Chapter number	2	3	4	5
Related topic from Section 1.3	language dynamics (Section 1.3.4) and language as capacity (Section 1.3.3)	diversity measurement (Section 1.3.2)	the socio-economic impacts of ethnolinguistic diversity/language situation (Section 1.3.1)	the socio-economic impacts of ethnolinguistic diversity/language situation (Section 1.3.1)
Related goal from Section 1.4.1	Goal 3	Goal 1	Goal 2	Goal 2
Related question from Section 1.4.1	Question 5	Question 1 Question 2	Question 4	Question 3
Main results	History explains the differences in the current status of languages in Sub-Saharan Africa. Missionary activities have contributed to higher polarization among linguistic groups. Colonial policies and geography have influenced language status patterns indirectly.	The ICP can be a useful measure in studies concerned with the development effects of diversity. The ICP is a multifaceted linguistic indicator that can be used for language policy planning and evaluation purposes, for instance.	Second languages/multilingualism has positive societal effects. People living in regions with higher average ICP tend to trust unknown people more.	Second languages have identification role. People who speak more than two languages tend to feel more national than ethnic. The ICP is not found to have a similar function on the whole sample, only in former French colonies.
Investigated dimension of the language situation	language status	multilingualism	ICP	ICP and the number of spoken languages
Level of analysis	language/ ethnolinguistic group	country	individual	
Number of observations	389	20	21642 individuals (nested across 229 ethnic groups)	19340 individuals (nested across 165 ethnic groups)
Linguistic data	Ethnologue (Lewis et al. 2014)	Afrobarometer, other surveys (national censuses, Demographic and Health Surveys etc.), Albaugh (2014)	Afrobarometer	
Method	OLS, instrumental variable technique, counterfactual analysis	descriptive statistics, graphical representation of the ICP	multilevel modeling (level 1: individuals, level 2: regions)	multilevel modeling (level 1: individuals, level 2: ethnic groups)

1.5 Future perspectives

There are two main directions for future research. One is a deeper analysis of the impact of state formation on the language situation in the long run; and the other one is to assess the efficiency of language policy in shaping language dynamics and improving the standards of living of minority and endangered groups. While the former is linked with the political economic literature on the role of early state forming, the latter has direct policy relevance.

As we discussed in Section 1.2, the historical forming of centralized states is an obvious candidate to explain the difference in language patterns across the globe. Although there is empirical evidence that younger states exhibit higher ethnolinguistic fragmentation (Kaufman 2015, Ahlerup and Olsson 2012), we still do not know much about the effects of nation building on other dimensions of the language situation such as the spread of second languages. Moreover, characteristics other than the age of a state such as the spatial and temporal stability and the state's origin (endogenously emerged or externally introduced) are also expected to matter. Such a project would help us explain not only the current diversities in linguistic patterns around the world but is expected to give some indication on future trends. Is monolingualism the ultimate fate of nation states or is bi- and multilingualism also sustainable? What can we expect in the developed world where immigration is leading to increased linguistic and cultural diversity? Or, could English as a global lingua franca crowd out local languages in the era of globalization?

The other potential research direction is to analyze the possibilities of language policy in shaping language dynamics and in reducing poverty and increasing social inclusion in African societies. As discussed in Section 1.3.3, minority language speakers are often excluded from education and cannot participate in political decision making due to language difficulties. Since they constitute only a small fraction of the society, their interests are less likely to be represented. Being able to speak up is an essential requirement of democracy, which is widely accepted as the desired form of government. By language policy not only laws regarding the official languages and the recognized mediums of education are meant, but all other rules and actions including language standardization and revitalization programs that directly influence individual language choices at the micro-level and language dynamics at the macro-level. The final aim of such a project would be to provide an empirical assessment of the outcome of language policy actions in Sub-Saharan Africa.

2 The historical determinants of language status in Sub-Saharan Africa[14]

Abstract

Languages are one of the most naturally evolving human institutions. Although the status of languages is closely associated with the well-being of their speakers in multilingual societies, this issue gains only a marginal attention in economics and development studies. This chapter aims to reveal the long-term determinants of the status of languages in Sub-Saharan Africa, one of the most linguistically fragmented areas of the world. Based on economic, anthropological and historical studies, we identify the following factors that are likely to have long-term effect on the current status of African languages: geography, precolonial contact with Europeans and the Arabs (Islam), precolonial development of indigenous societies, Christian missions and colonial policies. The main data sources are the Ethnologue, the Joshua Project, Murdock's Ethnographic Atlas, Roome's map on the location of missions, various sources on the first Bible translations in African languages, and geographical data available online in shapefile and raster format. Using OLS and IV estimation techniques, we find that indigenous groups with relatively high socio-economic development before the European dominance, early Bible translation and relatively large share within current country borders are less likely to have their language in an endangered state today. Geographical variables and the nature of colonial policy seem to affect current language status indirectly through their impact on socio-economic development and missionary activities. The counterfactual analysis suggests that the contact with Europeans contributed to higher polarization in terms of language status.

Keywords: Sub-Saharan Africa, language status, precolonial socio-economic development, European colonization, Christian missions, counterfactual analysis

---

[14] This study is published as CGEH Working Paper No. 66, online available at http://www.cgeh.nl/sites/default/files/WorkingPapers/cgehwp66_buzasi.pdf. The author is grateful to Jan Luiten van Zanden (Utrecht University), Maarten Mous (Leiden University), and the participants at the Utrecht Graduate Seminar (Utrecht, The Netherlands, 2015), the FRESH Meeting (Trinity College, Dublin, Ireland, 2014) and the Oxford Economic and Social History Graduate Workshop (University of Oxford, UK, 2014) for useful comments. A different version of this chapter written together with Peter Foldvari is going to be presented at the European Economic Association Conference (Mannheim, Germany, August 2015) and the European Historical Economics Society Congress (University of Pisa, Italy, September 2015).

2.1 Introduction

Sub-Saharan Africa is the most underdeveloped region of the world. According to the World Bank (2013), the fifteen poorest countries measured with per capita GDP (in current USD) are located in this region. Moreover, Sub-Saharan Africa exhibits the lowest average Human Development Index (United Nations 2014) and the perceived level of corruption is also strikingly high there (Transparency International 2014).

The existing literature provides various explanations for this extremely poor performance. Ethnolinguistic fragmentation (the probability that two randomly selected people in the society belong to different ethnolinguistic groups) has been found to undermine economic growth by increasing corruption (Mauro 1995) and the probability of insufficient public good provision (Easterly and Levine 1997). Another strand of the literature highlights the role of history in explaining Africa's exceptional (under)development path. Acemoglu et al. (2001) argue that where they could not settle due to unfavorable disease environment, European colonizers established extractive institutions which persisted to the present. Consequently, countries with higher settler mortality during the colonial era tend to have, for instance, worse property rights and weaker rule of law today. The proportion of white settlers is found to be a good predictor of contemporary institutional quality (Angeles and Neanidis 2015, Wietzke 2015). The nationality of colonizer also matters. Colonial policies implemented by different nations have different long-term economic and social impacts (Bertocchi and Canova 2002, Brown 2000, Grier 1999). Missionary activities, which include education (literacy, agricultural and technological knowledge) and medical care, have been established to be responsible for development differences within Africa. In areas where colonial rules did not restrict the activities of the different denominations disproportionately, competition between missionaries resulted in better schooling outcomes (Gallego and Woodberry 2010). And finally, the precolonial diversities in societal organization across ethnic groups seem to explain contemporary between- and within-country development differences (Michalopoulos and Papaioannou 2015 and 2013, Bandyopadhyay and Green 2012).

What is quite striking is that although the regulation of language use in public administration and schooling was a crucial element of colonial and postcolonial policies and it was missionaries who initiated the first linguistic works on the continent, language-related issues, other than ethnolinguistic diversity, gain only marginal attention in economics, economic history, and development studies. Languages, however, determine individual well-being and social welfare through various channels. In economics, language (in its abstract sense) is viewed as an institution which makes information flow cheaper and easier, thus facilitates communication and cooperation between individuals and groups (Smith 2010, Ostrom 2000). Better and less expensive communication and cooperation at the level of individuals can result in better organization at the macro or societal level. Due to lower

translation costs and benefits from the economies of scale, markets are expected to work more efficiently in linguistically homogenous countries (Martinez-Zarzoso 2003, Egger 2002). Language learning is a form of human capital accumulation which is supposed to increase personal earnings (Grenier 1982). However, certain languages have higher labor market value than others (Levinsohn 2007). Language as an observable cultural marker serves not only as a basis for social fragmentation, stressed by the ethnolinguistic diversity literature, but also as a basis for discrimination on the labor market (Pendakur and Pendakur 2002). Moreover, the speakers of the officially not recognized languages are likely to perform worse in schools, know less about infection and disease prevention, and less able to participate in political decision making (Harbert et al. 2009, Batibo 2005). Overall, linguistically fragmented regions are more likely to suffer from severe poverty (Romaine 2009).

This chapter gives insight on how history has shaped the linguistic situation, or more precisely, one of its aspects, in Sub-Saharan Africa. The study has several novelties. First, instead of the ethnolinguistic fragmentation, which has already been extensively researched, the status of languages is put into the center of investigation. We analyze the role of the precolonial socio-economic development level of indigenous groups, European colonization and missionary activities in explaining why certain languages are officially recognized, standardized and have higher social prestige than others. Second, the units of analysis are the linguistic groups in this chapter. Focusing on linguistic groups instead of countries makes it possible to understand within-country social differences (explained in more details in the next paragraph). Development studies explaining the impacts of ethnolinguistic fragmentation usually compare countries. Third, unlike historical and sociolinguistic studies that conduct case studies limited to certain languages or countries, our empirical analysis attempts to arrive at more generalizable results by incorporating almost 400 languages located in 47 Sub-Saharan African countries.

The results are broadly linked to the following research fields. Since the status of languages is closely associated with the well-being and capacities of their speakers, revealing the long-term determinants of the status of languages contributes to our understanding on the historical roots of socio-economic inequalities. Moreover, since languages are generally concentrated on fixed geographical territories, the factors that are responsible for the differences in language status are, at the same time, expected to explain regional (within-country) development differences to a certain extent. Finally, although it is highlighted above that this chapter focuses on an aspect of the linguistic situation other than ethnolinguistic diversity, our results are indirectly related to that research direction as well. While anthropological and economic studies have revealed that certain geographical and climatic factors (proximity to the Equator, good land quality, and low variability in rainfall) imply higher ethnolinguistic diversity (Kaufman 2015, Green 2013, Michalopoulos 2012, Cashdan 2001, Nettle 1999), the role of history is less known in this respect. Sociolinguistics provide evidence that if the expected benefits from identifying with a language are not high enough, people are

likely to abandon that language and acquire a new one with higher economic or social values (Mesthrie et al. 2009: 248-251, Batibo 2005: 93-94). Thus, since the status of languages is positively associated with the rewards from using that language, historical factors that influence the status of languages also determine which languages are maintained and abandoned. In the end, these individual decisions are expected to influence the size and the number of linguistic groups within a territory, i. e. linguistic diversity.

The chapter proceeds as follows. The next section provides an overview on the most important historical factors that have shaped the linguistic situation of Africa in the past centuries. Also, we derive five hypotheses that are going to be tested in the empirical section. Section 3 describes our data. The empirical analysis and the discussion of the results are presented in Section 4. The last section concludes.

2.2 Historical background

This section provides an overview on the main periods of Africa's linguistic history which serves as the theoretical framework for the empirical analysis in the next section. We discuss the socio-economic development of indigenous societies, missionary activities and European colonization from the aspect of their linguistic impacts. Although contact with Europeans had been taken place continuously via trading and missions from the 15th century onward, the terms 'European colonization' and 'colonial era' in this chapter refer to the period after the Berlin Conference in 1884-85 when the scramble for Africa began and the geographical boundaries were officially set up.

2.2.1 Linguistic situation before the colonial era

Before the European colonization, language dynamics in Africa were predominantly driven by internal factors such as geographical and ecological conditions, invasion, and assimilation; and external factors such as the Arab conquest beginning in the 7th century and the establishment of Christian missions from the 15th century onward. Trade, which was determined by both internal and external factors, is special in this listing.

Empirical studies have shown that ethnolinguistic diversity is positively associated with closeness to the Equator (Cashdan 2001), longer mean growing season (Nettle 1999), and low variability in land quality, rainfall, temperature and elevation (Kaufman 2015, Green 2013, Michalopoulos 2012). The most commonly accepted theory to explain this relationship is that ethnolinguistic diversity is higher in areas where geographical and climatic conditions make self-sustaining relatively easy, hence large scale cooperation and common actions between people or groups are not encouraged or necessary (Nettle 1999).

Although the climatic conditions of the Sahel region did not favor settled agriculture, the presence of rivers, the lack of physical barriers and the existence of salt and gold as valuable products of that time facilitated long-distance trade which became the basis of centralized kingdoms after the 8th century (Mansour 1993). The potential reward from trade served as the base for centralized states in other parts of the continent as well (Fenske 2014, Bates 1983). Wars and the expansion of these precolonial empires often resulted in major social, political and linguistic changes of the invaded societies. The enlargement of the Mali Empire, for instance, from the area of the current administrative region of Segou in the 13th century toward the oceanic coast in the West in the 16th century resulted in the massive spread of the Manding language (Mansour 1993).

Anthropological and historical documents argue that ethnic identity as understood by the modern Western world did not exist as a reality in Africa before the arrival of Europeans and 'group membership' was a highly fluid concept (van den Bersselaar 1997, Harris 1988). Ecological and economic shocks (famine and drought) often resulted in the displacement of people to new areas, reconfiguration of existing group ties and dependency of the dominant local groups (Stock 2012 p. 86.) which did not leave language use patterns intact. Murdock (1959) lists several precolonial examples on how certain groups shifted to the languages of their economically more developed neighbors. For instance, the Sanye, a hunting tribe in current Kenya, took over the language of the neighboring Bararetta Galla (Oromo) from whom they adopted the rudiments of animal husbandry. Other examples include the Beni-Amer (current Eritrea) that adopted Tigre and the Shabelle (current Somalia) who completely acculturated to the dominant Somali.

The first missionaries arriving in the mid-15th century with the main aim to explore Africa were predominantly Catholics sponsored by the Portuguese. Protestant missions started to operate only in the mid-18th century (Asafo 1997, Welmers 1974). Beyond spreading the Word of God and translating the Bible to local languages (discussed later), they provided agricultural and technical education.

Although trade was essential at the local level as well, it is trading between neighboring and often culturally distinct groups and international trade (the export of slaves, gold and ivory to Europe and America and the import of clothes) that are expected to have had more important linguistic consequences.[15] First, intergroup and interregional trading fostered the emergence of indigenous lingua francas (Heine 1970) such as Hausa (Nigeria, Niger), Songhay and Manding languages (Mali) and Swahili (Tanzania, Kenya, Uganda) which still have high esteem and important role in interethnic communication in contemporary African societies (Lewis et al. 2014, Mansour 1993). Second, due to the strong Muslim Arabic conquest associated with

[15] A good review on the geographical extension and the main products of local, intergroup and long-distant trade in Tropical Africa is found for instance in Konczacki and Konczacki (1977).

high political and religious centralization and control over the trans-Saharan and Red Sea trade routes, Arabic, a non-indigenous language, was also used as a lingua franca. Third, Africa provides several examples that lingua francas in the long run had spread to the expense of local languages and became primary languages for certain groups. The founders of the Niger-Gambia trade axis were the Soninke people who adopted Manding in the heyday of the Mali Empire (Mansour 1993). The speakers of Tumzabt, a threatened language in Algeria, also use Algerian Arabic as primary language (Lewis et al. 2014). Fourth, trade with Europeans accompanied with agricultural and technical education, and medical services offered by Christian missionaries (Welbourn 1971) contributed to positive attitude toward European languages. Several historical sources document the demand of indigenous people for European languages and missionary education before the establishment of the colonial administration. In a letter to a trading firm with the goal to get missionaries around 1850, the chiefs of Bonny (Nigeria) wrote: 'We expect that those gentlemen to be sent to us shall be capable of instructing our young people in the English language' (Ajayi 1965 p. 56).

Written form of languages was far from common and highly concentrated to the North, the Horn of Africa and the Eastern coastal region. Although a few languages developed their own scripts (for instance the Ge'ez of Amharic and Oromo in Ethiopia and the Tifinagh of Berber languages in North Africa), the majority of tribes or kingdoms that had written languages before the arrival of Christian missions were those participating in the trans-Saharan or the Indian (slave) trade and strongly influenced by the Islamic (Arabic) culture. Hausa (Nigeria, Niger, Benin), Tamasheq (Mali), Kanuri (Nigeria), Wolof (Senegal), Soninke (Mali, Senegal), Fulfulde (Nigeria), Swahili (Tanzania, Kenya, Uganda) and Songhai (Mali, Niger) languages used a version of the Arabic script (ajami) that was adopted to the writing of non-Arabic languages (Albaugh 2014, p. 23, Warren-Rothlin 2009, Adegbija 1994). The earliest linguistic works (Bible translation, dictionaries and the development of Latin script for African languages) by the Christian missions were related to languages already existing in written form. According to Groves (1964), the first Bible portion (Psalms) translation on the African continent was compiled in Abyssinia in 1513. Several Coptic translations appeared in 1663, 1786, and 1811. Bible portions were translated to Amharic in 1824, Oromo in 1839, Swahili in 1847, Hausa and Kanuri in 1853, and Wolof in 1873. Missionary linguistic works related to languages along the coast and located around early trading ports were implemented also relatively early. Although Fanti (one of the Akan dialects in current Ghana) is the only non-Sahel region language in which Bible portions were available before 1800, several parts of the Script were translated for instance to Ga (Ghana) in 1805, Susu (Guinea and Sierra Leone) in 1816, Nama (Namibia, South Africa) in 1831, Xhosa (South Africa) in 1833, Grebo (Liberia) in 1839, and Zulu (South Africa) in 1846.

Numerous historical sources inform us on the nature and intensity of the relationship between European missions and local people, or more precisely, kings, chiefs and the elite in the 19th century. In 1823, Radama I (1788-1828), the Merina

king in Madagascar, decided to modernize his country to keep Europeans at a distance. He proclaimed the Latin script of Malagasy introduced by Reverend Jones (London Missionary Society) to be the official alphabet of the kingdom and allowed the LMS to establish schools, develop teaching materials and promote literacy. However, schools concentrated on the central highland and were available only to the Merina nobility (Steinhauer 2005 pp. 78-80). Crowder (1968 p. 9.) points out that the chiefs in Sierra Leone were asking for missionary schools twenty years before the establishment of the Protectorate in 1896. Ajayi (1965) reports an excellent relationship between the Efik king (Eyo) and the mission under Hope Waddell (1804-1895) arriving in Calabar (Nigeria) in 1846. To adapt to the changes in international commerce, Eyo wished to support the transition from a former slave-based economy to palm oil production through education provided by missionaries.

2.2.2 The linguistic effects of European colonization

The 'scramble for Africa' beginning in the 1880s not only opened up a new era in African economic, social and political history, but significantly influenced the linguistic landscape of the continent.

Colonial boundaries settled in 1884-85 at the Berlin Conference are commonly accepted as a result of arbitrary decisions (Hargreaves 1985, Englebert et al. 2002) without taking natural geographical circumstances, original social organizations and indigenous group distribution into account. An important consequence of the fixed colonial borders is that they anchored or froze the previously dynamic economic and political relationship between indigenous groups and determined which groups were to compete with each other for the colonial administrative positions, education and other missionary services. Dividing some groups between countries might have resulted in the weakening of certain groups and the strengthening of others. As Posner (2004b) shows, the current political relevance of the Chewa and Tumbuka people (both divided between Zambia and Malawi) is highly dependent on their relative size within the current country borders.

Although, as it is discussed above, missionary activities were significant from the 15th centuries, the massive penetration of the inland regions of Africa and organized linguistic works started only in the late 19th and early 20th century. Missionaries were dedicated to language standardization and development through establishing the written form of certain African languages (Latin script), compiling dictionaries, grammar books and orthographies that served education, literacy promotion and Bible translation purposes (see for instance Peterson (1997) on the process of creating the Gikuyu dictionaries in Kenya or Doke's report (1931) on the creation of a single Shona orthography).

Due to resource scarcity and other practical reasons, missionaries often had to decide which languages or dialects should be developed. For instance, Hausa was selected for standardization for its beneficial properties. Since it already existed in

written form and due to its function as a lingua franca in trade, the cost and benefit (reaching several tribes with one language) ratio of its development, standardization and promotion was expected to be smaller compared to its non-written and locally concentrated counterparts (Adegbija 1994). However, there were other decisive linguistic characteristics. In relation to Southern Africa, Gilmour (2007) highlights that some languages (such as Xhosa and Tswana) were found to be sufficient and promoted for religious translation, others were considered ineligible (lacking of crucial religious concepts) and inefficient (certain terms needed to be expressed with circumscription), thus became discouraged (such as Khoi and San languages). In certain cases a single dialect was chosen as the basis for standardization (such as Ki-Unguja (Zanzibar) dialect of Swahili over the Mombasa dialect), in other cases the unified orthography was based on various dialects (the Shona language was a distilled variety from the Zezuru, Korekore, Karanga, Manyika and Ndau dialects) (Chimhundu 1992, Ansre 1974, Whiteley 1956).

However, missionary activities cannot be viewed isolated from the system of the colonial rules in which they operated. Missions were not only the receptive subjects of the colonial administration with the exclusive aim of spreading the Word of God but they represented significant economic and political power. They contributed to setting up economic and social conditions (such as creating new markets) that made colonial utilization profitable (through training masses and local administrators) (Fabian 1983). An old Gikuyu (ethnic group in Kenya) proverb 'There is no difference between a missionary and a settler' (Oliver 1952) illustrates that Christian mission was perceived by local people as the representative of the colonial power.

Each element of the colonial practice (including language and education policy and the regulation of missionary activities) implemented by the different European empires were designed to match their overall underlying philosophies. The French and Portuguese considered Africans as people that need to be civilized (Conklin 1997) and to be assimilated into the broad metropolitan community (Betts 2005, Bokamba 1991, Spencer 1974) in the long run. The exclusive use of French and Portuguese in the colonial administration and at all levels of education was seen as a main tool to achieve these goals. Local languages were not only not promoted or ignored but actively discouraged (Bokamba 1991, Spencer 1974). In contrast, the British acknowledged local circumstances to a greater extent. Local languages were extensively studied, documented and standardized and some of them (mostly dominant local vernaculars) were used as the language of education in primary schooling. Mother tongue education was seen as a tool to support the success of English learning in the long run (Atkinson 1987, Ward 1940). However, English served as a language of higher education and the colonial administration (Berman 1975).

Education policy (beyond the rules regarding the language of instruction) was also designed in line with the general colonial vision. Although the French government established numerous public schools and allowed limited (mostly Catholic) missionary activities, Hailey (1945 p. 1261) points out that missionary activity remained sufficient

and the proportion of students enrolled in public schools in French West Africa did not exceed 15%. British preferred to 'outsource' most of their education to missions irrespective of their denomination (Frankema 2012) and all religious groups could apply for grants from the state (Gallego and Woodberry 2010). The French colonial education system is often described as highly selective and 'elitist' which provided disproportionate access to the French type education, and consequently, to prestigious state positions for certain ethnic groups (Blanton et al. 2001 p. 478). Compared to the French, the British enrolled more students into the education system and invested higher percentage of their budget in schooling (Frankema 2012, White 1996). However, due to high selectivity, the French system resulted in higher per capita education expenses (White 1996).

European colonization induced several changes in the economic and political organization of traditional societies. Based on a fieldwork conducted between 1949 and 1952, Middleton (1971) provides an overview on the consequences of European colonization in the Lugbara (Uganda and the Democratic Republic of Congo) society. Due to railway and the introduction of the tobacco as a cash crop during the era of the British administration (1894-1962), from a subsistence economy, the Lugbara people transitioned to a peasant economy with larger and continuous settlements, labor work and institutionalized markets. Instead of hereditary succession, chiefs that ruled certain geographical units (counties, sub-counties and parishes) with fixed borders were appointed by the government. Although older generations considered this period as the destruction of traditional values, younger people saw its economic advantages.

From the aspect of this study, the most important consequence of the above discussed policies is that they changed the costs and benefits of maintaining local languages which determine language dynamics (language maintenance and decline) in the long run (Harbert et al. 2009). Since they were attributed with political power and prestigious socio-economic status, both French and English enjoyed favorable attitudes (Bunyi 1999), which is expected to have increased the incentive to learn them. Yet, this encouragement could have been counterbalanced in former French colonies by two factors. First, since ethnic groups experienced unequal chances to gain state positions, individual motivation to learn French might have remained low among the members of the less preferred ethnic groups. Second, if lower proficiency in mother tongue is associated with lower success in second language acquisition as suggested by the socio-linguistic literature, the French system with the language of the colonizer as the exclusive mean of instruction is expected to work less efficiently compared to the British system. Local languages selected for education purposes also gained considerable esteem (for instance the Efik in Nigeria) (Adegbija 1994).

2.2.3 The postcolonial era

Since independence did not result in essential changes in the nature of language policies[16], the aforementioned colonial practices have several long-term linguistic and language-related social consequences. Due to the ignorance and discouragement of local languages, former French West Africa still had the least developed lingua francas in the world in the beginning of the 1990s (Bokamba 1991). While the share of the population with a written language reached 76%, 79% and 80% in 1950 in British, Belgian and Portuguese colonies respectively, the corresponding share was only 58% on the French territories (Albaugh 2014 p. 27 and p. 70). Recent empirical studies show that former British colonies still exhibit higher average levels of education and literacy (Cogneau 2003, Brown 2000) which is partly the consequence of higher competition between the Protestant and Catholic missions (Gallego and Woodberry 2010).

The most remarkable novelty of the past two decades in linguistic terms has been the increased focus on endangered and minority languages. Having acknowledged the severity of social and ecological issues related to language decline (identity and culture loss, difficulty in accessing education and health services, the high possibility to get excluded from political decision making, and the interconnectedness between biodiversity, linguistic diversity, and poverty) a number of international organizations and research groups have committed to elaborate policy guidelines to support language revitalization programs that meet global, regional or local needs.[17] In parallel to this worldwide phenomenon, the majority of former French colonies have recently moved toward the recognition of local languages in education (Albaugh 2014).

2.2.4 Hypotheses

Based on the literature discussed above, we derive the following five hypotheses in relation to the current status of Sub-Saharan African languages. In this study, language status is understood as a complex concept which is defined in Section 2.3.1 in more

[16] The exclusive use of French was maintained after independence in most countries (except for Northern African countries and Madagascar) to avoid tribal conflicts and providing access to global development by using a language of wider communication (Bokamba and Tlou 1977).

[17] In the past decades language decline and the loss of linguistic diversity have been acknowledged a serious problem affecting various areas of the globe. According to the latest edition of the Ethnologue, 34 percent of the currently known languages are threatened (likely to die out in the short run) and about 370 languages have died out (lost all native speakers) since the 1950s. The UNESCO launched its Endangered Languages Program in the early 1990s and beyond helping policy makers through professional advices and meetings it publishes its results as the UNESCO Atlas of Endangered Languages (Moseley 2010). The SIL International (http://www.sil.org), the largest Christian non-profit organization, is engaged in studying and documenting languages of the world in order to promote literacy, support Bible translation and language development policies. Their results are published in a comprehensive catalog titled Ethnologue: Languages of the world that serves as reference work in several disciplines addressing language related issues. The Catalogue of Endangered Languages (ELCat, http://www.endangeredlanguages.com/), a joint project of The LINGUIST List and the University of Hawai'i Mănoa, is a recent attempt with similar goals as the above mentioned two examples.

detail. For now, the following 'definition' is sufficient. Languages with official recognition in public administration and/or education, with established orthography and high social esteem are considered to have higher status compared to those that are officially not recognized, unwritten and not favored by the society.

First, we expect that the languages of those groups that were more developed in social and economic terms before the European dominance have higher status today (Hypothesis 1). This argument is supported by the sociolinguistic and economic literature. Batibo (2005 pp. 93-94) argues that language shift takes place when a weaker linguistic group comes into contact with another language that has higher demographic, economic, cultural or political value; the speakers of the weaker language want to identify with the other group in order to share in the benefits attached to that language. This is in line with 'identity economics' (Bodenhorn and Ruebeck 2003) which views identity as a fluid concept that is not 'determined by nature' but dependent on individual choices. If the benefits from acquiring a new identity are higher than the costs of abandoning the existing one, people are likely to give this latter up and choose the more advantageous strategy.

Second, size matters. It is commonly accepted that larger communities have higher chances to survive and spread widely since they execute demographic pressure on smaller languages (Batibo 2005). Also, more populous ethnic groups have more bargaining power to secure a higher legal status and financial sources for their languages. More speakers may also give rise to a kind of economies of scale, since with the same investment in learning a language more possible partners can be reached (or, alternatively, this can be seen as a network externality). As a result, we expect that the higher the share of a linguistic group within a country is the higher the language status should be (Hypothesis 2).

Third, we also expect that languages that were standardized earlier have higher current status (Hypothesis 3). As the historical evidence suggests, due to scarce resources, missionaries did not develop each language equally (thereby they optimized their use of scarce resources). High initial costs of language development encouraged missionaries to focus on a few selected languages instead of many, while the possible increasing returns made it profitable to focus on those languages which had already achieved a higher degree of development. They also promoted the use of such preferred languages among other groups with different primary language.

Fourth, the number of missions on a territory is associated with the amount of human resources that can be utilized for linguistic works and the bargaining power within the colonial government. Thus, we expect that the higher concentration of missionary activities is associated with higher current language status (Hypothesis 4).

And finally, due to the promotion of local languages and the higher investment in primary education discussed in Section 2, we expect that the current status of African languages is higher in former British colonies (Hypothesis 5).

2.3 Data, variables and limitations

2.3.1 The dependent variable: language status

Similarly to most phenomena investigated by social and political scientists, the status of languages can be approached from various aspects. On the one hand, it is possible to measure status from one angle only (minimalist approach). By their legal status, languages can be officially recognized as the language of national or regional affairs and education or officially not recognized in any socio-economic domain. Languages without established orthography and available written material are usually considered inferior to standardized languages. Moreover, due to the number of speakers and their traditional role in trade for instance, some languages have higher social prestige (e.g. lingua francas) than others. On the other hand, the status of languages can be understood as a complex concept determined by various factors including those just mentioned above (maximalist approach).

This work relies on the maximalist approach for the following reasons. First, since we aim to link our results to the broader development and ethnolinguistic diversity literature, we have to take into account that the well-being of the speakers and the linguistic choices of individuals are determined by linguistic factors other than their official recognition. For instance, although the Dioula language in Burkina Faso is not recognized officially, as the traditional trade language, it is widely used for interethnic communication and has high prestige. The second reason is that there is an available appropriate measure.

In this chapter, language status is measured with the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Lewis and Simons 2010) from Ethnologue (Lewis et al. 2014), which summarizes numerous potential aspects of the language status in a single indicator. The original purpose of the EGIDS is to provide an appropriate tool that helps evaluate the severity of language endangerment and to determine the nature of actions that are required to overcome language decline (Footnote 14). The vitality (or endangerment) of languages is evaluated along five dimensions. Languages are considered less endangered if they are actively used as first language (sociolinguistic abbreviation is L1) and vehicular or lingua franca (key concept #1: identity function), officially recognized at the national or regional level (key concept #2: official recognition), transmitted uninterruptedly from generation to generation as first language (key concept #3: transmission), taught and/or used in the education system (key concept #4: literacy status), and younger generations speak it actively as L1 (key concept #5: youngest generation). According to these concepts, a 13-level scale is derived (see Table A2.1) where higher values reflect more disruption, i. e. higher chance to die out in the short run. The scale levels are hierarchical in nature, which makes it possible to use it as an ordered (ordinal) dependent variable in the empirical models estimated in the next section. With only one exception, the scale assumes that each stronger level of vitality entails the characteristics of the levels

below. The one exception to this principle is level 3 (wider communication) where the vehicularity of languages is counted as being more important than the existence of an orthography and the use of language in education. Some languages that are widely used for intergroup communication are not used in education and have no written materials. Were these languages to lose that vehicularity, they would drop directly to level 6a. The main benefit of the aforementioned hierarchical structure is that if one of the underlying aspects changes, the status of the language also changes by holding the remaining aspects constant.

Compared to other existing language endangerment assessment methods (the Graded Intergenerational Disruption Scale (GIDS, Fishman 1991), the UNESCO's language endangerment classification scheme (UNESCO 2003, Moseley 2010), and the language status measures applied in the previous editions of Ethnologue), the EGIDS has various advantages which make it the best available language status measure for the purpose of this study. Unlike the UNESCO classification which handles safe languages as a single category and the GIDS by Fishman (1991) which does not provide an adequate description of the possible language statuses at the lower end of the scale, the EGIDS classifies languages into 7 non-endangered (from international (0) to vigorous (6a)) and 6 endangered (from threatened (6b) to extinct (10)) categories. Another benefit of the EGIDS compared to GIDS is that endangerment is not only dependent on the characteristics of the language itself (internal or micro factors such as intergenerational transmission) but on external or macro-level factors such as the social attitude, government support, and the institutional context. As Lewis and Simons (2010) argue, although the nature of intergenerational transmission is an eligible factor to distinguish between the different level of endangerment, at the upper end of the scale the level of institutionalization might be a more appropriate criteria to distinguish between the more developed levels.

Among the other possible classification schemes, the EGIDS is the most dynamic: it does not only assign high importance to intergenerational transmission as the GIDS does, but the final categories are designed to indicate whether a language is on the way to language shift or language development.

And finally, due to its hierarchical and refined structure, the EGIDS can easily be collapsed into only one aspect of the language status. This property is useful in two situations. First, if the research question requires information only on the official status, for instance, the researcher can consider all languages with EGIDS score above 5 as recognized languages. Similarly, all languages above EGIDS 6a have established orthographies. Or, all languages with EGIDS score above 6b can be seen as safe, while the remaining ones as threatened languages. Second, complex social, economic and political indicators (e.g. the Human Development Index[18] or the polityIV[19] which is

[18] http://hdr.undp.org/en/content/human-development-index-hdi [05 May 2015]
[19] http://www.systemicpeace.org/polityproject.html [05 May 2015]

designed to measure democracy) similar to EGIDS are often criticized that they are too vague or rely on arbitrarily selected components. For the beneficial characteristics discussed above, the EGIDS can be applied even in case the researcher prefers the minimalist approach.

2.3.2 Sample design

The basis of our dataset is a map by Murdock (1959) (available in shapefile format on Nathan Nunn's[20] website) that shows the geographical location of the indigenous ethnic groups (835 groups) in Africa and provides an appropriate framework to store and organize the data derived from various sources for an empirical analysis with individual languages (or linguistic groups) as the units of analysis (Figure A2.1). Although the map and most of our sources would make it possible to conduct an investigation covering the whole African continent, the availability of information on the social and economic characteristics of indigenous African societies enables only a more limited analysis.

Data on the dimensions of socio-economic development of local societies are derived from the Ethnographic Atlas (Murdock 1967) which contains anthropological data on 299 African groups. But, since we aim to control for the effects of colonial rules practiced by different European empires and the potential effect of borders, we account for the partitioning of groups between countries. Thus, for instance, the Kung group which is divided between Namibia and Botswana is included in our database twice. This procedure also makes it possible to account for the fact that the status of the same language often varies by country. For instance, the status of the Venda language is level 1 on the EGIDS in South Africa, while in Zimbabwe it is only level 5.

Some groups are excluded from the final dataset and the empirical analysis for certain reasons. Due to the presence of Arabic long before the European colonization, six groups located in Algeria, Tunisia, Libya, Western Sahara, Morocco, and Egypt (Ahaggaren, Barabra, Kabyle, Mzab, Riffian, and Siwans) are left out. Some other groups are ignored for more practical reasons. Certain groups represented in the Ethnographic Atlas are considered as only dialects and not separate languages in Ethnologue. In this case we followed two strategies. First, if the group in the Ethnographic Atlas is a dialect in Ethnologue and the related main language is not included in the Ethnographic Atlas (for instance Pondo as a dialect of Xhosa and Lovedu as a dialect of Pedi in South Africa, and Xhosa and Pedi themselves are not included in the Ethnographic Atlas), we use the information on the dialect group as a proxy for the main language. Second, if both the main language and its dialect(s) are represented in the Ethnographic Atlas (for instance Afikpo as a dialect of Igbo in Nigeria or the Luapula dialect of Bemba in Zambia), dialect groups are eliminated.

[20] http://scholar.harvard.edu/nunn/home. [12 Nov 2014]

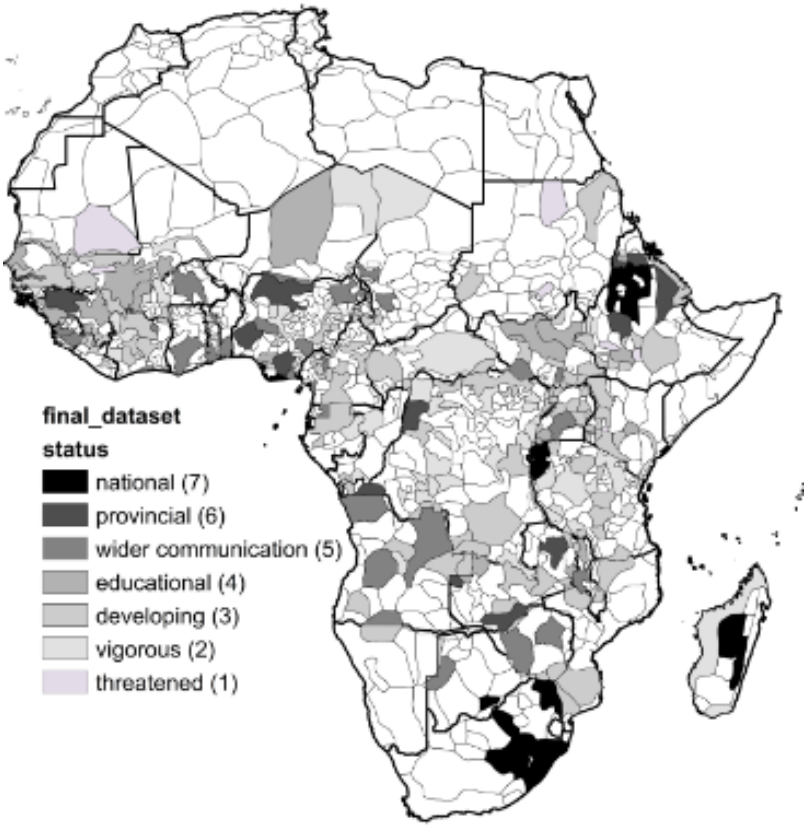
These aforementioned restrictions and the controlling for the country borders result in 389 Sub-Saharan African groups in the final dataset.

Then, the 389 groups from the Ethnographic Atlas are assigned with an EGIDS level of language status derived from Ethnologue discussed in the previous subsection. The distribution of languages per status level is presented in Table 2A.1. The distribution of languages across status categories in our sample is somewhat different from that found in Ethnologue regarding the whole continent. Instead of EGIDS level 6a (vigorous), EGIDS level 5 is the most populous category in our sample. Our sample does not include international, nearly extinct, dormant and extinct languages. Due to the small number of languages, level 6b, 7 and 8a are combined and understood as endangered or tongues with the lowest status. Moreover, since our analysis is concerned with the status of languages instead of their endangerment level, the scaling is reversed in order to ease the interpretation of the coefficients in the empirical analysis. Data availability is shown in Figure 2.1.

At this point, however, we have to remark that combining the Ethnographic Atlas that refers to ethno-cultural groups with Ethnologue which basic units are languages requires some compromises. Our strategy implicitly accepts the common simplification of development studies that ethnic and linguistic groups are identical. Empirical works relying on national censuses and other surveys often use linguistic data to proxy the ethnicity of respondents (Cheeseman and Ford 2007). Nevertheless, this is not necessarily true, since during the course of history some linguistic groups (mostly minority language groups) have taken over the language of another group as primary tongue or several groups belonging to the same ethnicity or cultural group speak different primary languages today (some examples are given in Section 2,2.1.). As a consequence, groups in the Ethnographic Atlas and languages in the Ethnologue are not identical: certain groups in the Ethnographic Atlas could be assigned with several languages. When this is the case, the language with the highest reversed EGIDS level is assigned to the Ethnographic Atlas group. For instance, the Bete people in Ivory Coast in the Ethnographic Atlas can be assigned with three languages in Ethnologue: Bete-Daloa and Bete-Guiberoua with EGIDS level 5 (reversed score is 3) and Bete-Gagnoa with EGIDS level 6a (reversed score is 2), which, according to our strategy, results in the final language status of 5 (reversed score is 3).[21]

[21] This chapter does not intend to evaluate the method according to which languages are decided to be distinctive or dialects in the Ethnologue. For instance, although it is said that Malagasy languages in Madagascar are mutually intelligible, the Ethnologue reports them as separate individual languages not dialects. Only the Merina (Malagasy, Plateau in Ethnologue) reaches the highest EGIDS level of 1.

Figure 2.1 Data availability



2.3.3 Socio-economic development of indigenous societies

In order to test Hypothesis 1, we need a variable to proxy the degree of socio-economic development of indigenous societies before the European dominance. By today's standards, a society is considered more developed if it exhibits higher GDP per capita and urbanization ratio, better quality institutions (less corruption, rule of law, democracy), higher average life expectancy and education level etc. Although, due to data unavailability and conceptual challenges, the development of precolonial African societies cannot be measured along these dimensions, recent studies have shown that certain traits of precolonial societies strongly correlate with current development measures. Empirical studies utilizing anthropological data from the Ethnographic Atlas by Murdock (1967) and the Corrected Ethnographic Atlas by Gray (1999) find that the variation in the precolonial centralization level of ethnic groups explains regional- and country-level differences in economic development (proxied with satellite images of light density at night) (Michalopoulos and Papaioannou 2013), public good provision and the quality of contemporary institutions (Gennaioli and Rainer 2007) across the African continent. Moreover, Bandyophadyay and Green (2012) show that precolonial centralization is highly correlated with several measures of development within Uganda, which point to the persistence of wealth and poverty from precolonial times to present.

Although it also utilizes information from the Ethnographic Atlas (Murdock 1967), contrary to previous studies listed above, our study is not limited to a single variable. While previous works were concerned with only one particular aspect of the social development of indigenous African groups, namely the existence of a centralized state (variable 33 in Gray (1999)), we consider development as a wider, multidimensional concept. We assume that the degree of the socio-economic development of indigenous societies[22] cannot be observed directly, but can be estimated from observable factors that are related to it. In this chapter, the generalized structural equation model technique (gsem command in Stata 13) is applied to estimate the latent (unobserved) socio-economic development variable from various (observed) data in the Ethnographic Atlas that are related to the complexity of societal and economic organization and can be measured on an ordinal scale. Selected variables including the intensity of agriculture (variable 28), jurisdictional hierarchy beyond the community level (the second digit of variable 32), class stratification (variable 67), succession of the office of local headmen (variable 73), and the presence of high god (variable 34)) are described in Table A2.2. All these variables are ordered and higher values represent higher complexity.

The underlying idea behind the measurement model is described in Figure 2. We assume that all observed variables of the socioeconomic development reflect a single underlying continuous latent variable. Hence, the observed indicators are all imperfect measurements of this underlying factor with independent errors (denoted by e). Consequently, their observed correlation is due to the common latent development factor only. Moreover, when more than one proxy of a single latent variable is available, using all available indicators to estimate the latent factor is always preferable to a single proxy approach. The inclusion of more variables reduces the measurement error and avoids the bias resulting from focusing on only a single aspect of the latent factor.

The model outlined in Figure 2.2 operates under very similar assumptions to a factor analysis with a single factor, but while traditional factor analysis is based upon the assumption that all observed variables are continuous, here we allow them to be ordinal. This leads to more efficient and theoretically more appealing estimates. The coefficients are reported in Table 2.1.

The number of observations in Table 2.1 might require some explanation. It is said above that the number of languages (or groups) in the final dataset which takes colonial borders into account is 389. Still, the socio-economic development of indigenous societies (Table 2.1) is estimated on a sample that takes each group into account only once. Then, each division of the partitioned groups located in different

[22] Since we can assume that the Ethnographic Atlas does not represent the precolonial characteristics of African societies purely (details provided in Section 2.4.2.) we name our variable as 'socio-economic development of indigenous societies' and refrain from the term 'precolonial development'.

countries is assigned with the same estimated score of the latent socio-economic development variable. There are two reasons for doing this. First, the development variable is aimed to grasp the socio-economic characteristics before colonialism when tribes were not divided. And second, including groups as many times as the number of countries they are assigned to by the colonial borders would give more weight to divided groups within the dataset, which would bias the estimation.

Figure 2.2 The measurement model of socioeconomic development

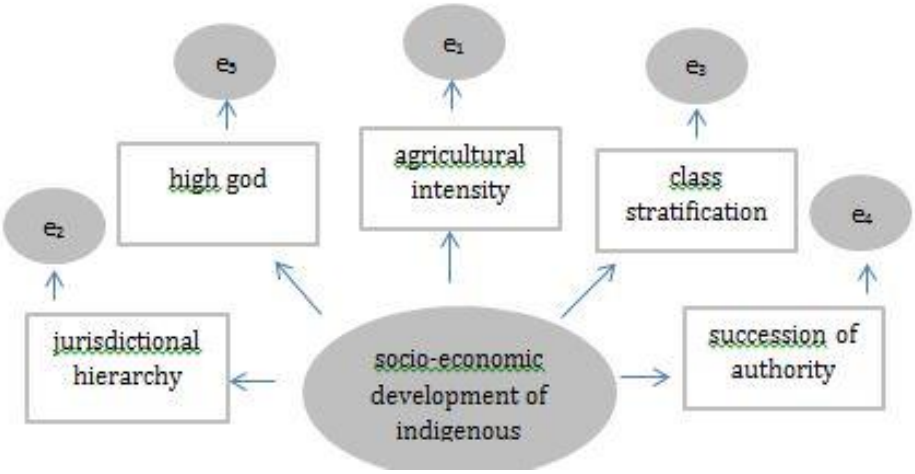


Table 2.1 Estimating the latent socioeconomic development of indigenous societies

	Model 1	Model 2	Model 3
	latent var: socio-economic development (dev 1)	latent var: socio-economic development (dev 2)	latent var: socio-economic development (dev 3)
agricultural intensity	**0.458*** (2.61)**	**0.611*** (3.10)**	**0.825*** (3.16)**
jurisdictional hierarchy (beyond community level)	**2.518** (2.42)**	**2.192*** (4.71)**	**2.049*** (4.27)**
class stratification	**2.509* (1.84)**	**2.962*** (2.87)**	**2.887*** (3.93)**
authority succession	-	**0.600*** (3.30)**	**0.413* (1.77)**
high god	-	-	**0.602** (2.35)**
number of observations	237	194	136
log-likelihood	-728.63856	-758.84932	-680.00877

Note: cuts are not reported, Z-stats in parentheses. . *, **, *** label significance at 10%, 5%, and 1% level respectively.

The three specifications in Table 2.1 are based on various sets of variables. It starts with a specification that leads to the highest number of observations (237 undivided groups) and continues with including additional components of development with less observations serving as robustness checks. We restrict the latent variable to have zero

mean and unit standard deviation. The latent development factors are strongly correlated (Table A2.3). While the procedure does not allow for a simple estimation of communality, Table A2.4 reports the rank correlation coefficients between the estimated latent development variables and their respective components. The latent socio-economic development variable is positively related to all included variables. According to the results, the Rundi (Burundi), the Ruanda (Rwanda, Uganda, and the Democratic Republic of Congo), the Amhara (Ethiopia), the Songhay (Burkina Faso and Mali), and the Oyo Yoruba (Nigeria) were the five most developed or complex societies (Table A2.5).

2.3.4 Population share, missionary activities and colonial policies

Ideally, testing Hypothesis 2 would require historical information on the share of ethnic groups within the colonial borders. Since these data are difficult to obtain, we can follow two, yet, imperfect strategies. First, we can utilize the area of each group in the Murdock map (1959) to proxy the share within the colonial borders. The benefit of this strategy is that it is based on historical information and the share of groups within the colonial borders effective before independence can be easily computed. The obvious drawback is that the appointed areas named after one certain group in the map might contain several other culturally-related groups and the area share without information on population density does not necessarily capture the population share. The second possibility, which is actually applied in this chapter, is to rely on the Joshua Project which provide data on the current share of ethnic groups within country borders (*variable: ln(pop share*)). Although this option makes it possible to work with numbers in relation to each known ethnic and linguistic group, the data mirror the contemporary situation which must be handled (Section 2.4.3).

In order to test Hypothesis 3, the year of language standardization is proxied with the year of the first Bible translation (either portions, New Testament or the complete Bible) in African languages represented in the sample. We deduct the year of the earliest reported translation in our three available sources (Groves 1964, Lewis et al. 2014, worldbibles.org) from 2014 (the publication date of the latest version of Ethnologue from which language status information is taken), which gives the final variable (*variable: bible age*) included in the empirical models. Groves (1964) reports the year of the first portion, New Testament and the complete Bible separately until the mid-1950s. The Ethnologue (Lewis et al. 2014) usually do not inform about portion translation, only on New Testament and the complete Script. The worldbibles.org dataset often reports unavailable information even in case of languages in which Bible is available according to the other two sources.

The location of Christian missionaries (Figure A2.2), which is required for testing Hypothesis 4, is obtained from Roome's (1924) map that has been extensively used in studies seeking the long-term development impacts of missionary activities in Africa. Although the source would make it possible, we do not distinguish between Catholic

and Protestant missions. We utilize the map from Murdock's Africa: Its people and their culture book (1959) digitalized by Nunn (2008) showing the geographical location and size of precolonial societies in the Ethnographic Atlas to calculate the number of missionaries per 100 square kilometers on their area (*variable: missions per 100 km²*). However, this variable is applied to proxy the European influence as discussed in Section 2.4.1.

The nature of colonial policies, which is the key variable in testing Hypothesis 5, including the use of languages in public administration and education is proxied with nationality of colonizer at independence (*variables: British, French, Belgian, Portuguese*) as provided by Bertocchi and Canova (2002).

2.3.5 Other variables: geography, climate and the spread of Islam

The empirical models discussed in the next section contain several control variables related to geography, climate and the spread of Islam. Geographic environment is controlled for with a number of variables. First, we control for the natural logarithm of the distance (in kilometer) of the geographic center of ethnic groups (calculated using ArcGIS 10.2.1) to the Equator (*variable: ln(distance to equator)*)[23], oceanic coasts (*variable: ln(distance to coast)*)[24] and inland waters (rivers and lakes)[25] (*variable: ln(distance to inland water)*). Second, each group is assigned with the areal mean and standard deviation of precipitation (in millimeter) (*variables: mean precipitation and std of precipitation*), mean temperature (in 0.1 Celsius degree) (*variables: mean temperature and std of temperature*) and altitude (in meter) (*variables: mean altitude and std of altitude*).[26]

As it is suggested by Bates (1983) and empirically tested by Fenske (2014), indigenous societies are likely to be more centralized where ecological diversity, hence the opportunity to profit from trade, is higher. Following Fenske's strategy based on the vegetation map by White (1983), the ecological diversity (*variable: ecological diversity*) per ethnic group is calculated. The measure (based on the well-known Herfindahl-index) can be explained as the probability that two randomly selected

[23] Shapefiles including tropical and polar circles, equator, and International Date Line is available at the website of Natural Earth. In this study we use the shapefile with large scale resolution (1:1000000) directly downloadable at http://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-geographic-lines/.

[24] Shapefile on the coastline of Africa is obtained from http://omap.africanmarineatlas.org/BASE/pages/coastline.htm.

[25] Shapefiles (medium scale resolution, 1:50000000) containing the main lakes and rivers in Africa are obtained from the Natural Earth database directly downloadable at http://www.naturalearthdata.com/downloads/50m-physical-vectors/50m-lakes-reservoirs/ and http://www.naturalearthdata.com/downloads/50m-physical-vectors/50m-rivers-lake-centerlines/ respectively.

[26] These climatic variables based on information for the period between 1950 to 2000 obtained at climate stations (presented here: http://www.worldclim.org/methods) are downloaded from the WorldClim – Global Climate Data website in ESRI grid (raster) format with the smallest resolution (10 arc-minutes). Direct link: http://www.worldclim.org/current.

points within the territory of an ethnic group belong to different major ecological zones.[27]

Retrieving historical data on the spread of Islam (*variable: islam*) before or during the colonial era in Africa is a challenge. But, since the share of Islam has been relatively stable in the past hundred years and Christianity could spread rather to the expense of traditional religions (Asafo 1997), we use the current share of each group affiliated with Islam as the proxy of early Muslim Arabic influence from the Joshua Project.[28]

Summary statistics of the variables introduced above are found in Table 2.2. Table A2.6 contains the data sources. Table 2.3 summarizes the expected signs of the main variables if the hypotheses in Section 2.2.4 hold.

Table 2.2 Descriptive statistics

variable name	#	mean	std. dev.	min	max
language status	389	3.234	1.371	1	7
socio-economic development 1	237	0	1	-1.619	2.439
socio-economic development 2	194	0	1	-1.62	2.531
socio-economic development 3	136	0	1	-1.828	2.478
ln(pop share)	388	-4.502	2.045	-10.677	-0.162
bible age	389	73.913	53.475	0	250
missions per 100 km^2	389	0.014	0.04	0	0.448
British	389	0.442	0.497	0	1
French	389	0.306	0.461	0	1
Belgian	389	0.100	0.301	0	1
Portuguese	389	0.049	0.216	0	1
ln(distance to equator)	389	6.664	1.033	0.693	8.169
ln(distance to coast)	389	6.041	1.136	1.099	7.469
ln(distance to inland water)	389	4.412	1.179	0	6.779
mean temperature	389	238.22	24.81	135	287
std of temperature	389	10.68	8.04	0	49
mean precipitation	389	56.89	77.21	0	371
std of precipitation	389	8.89	11.72	0	92
mean altitude	389	699.24	496.98	9	2234
std of altitude	389	164.81	148.19	2	819
ecological diversity	389	0.319	0.223	0	0.802
islam	389	0.254	0.359	0	1

[27] The original 81 ecological zones in White (1983) are collapsed into 18 major types. Fenske (2014) applies alternative measures to proxy the potential gains from trade (such as ecological polarization, distance to ecological border, or using the Food and Agricultural Organization's ecosystem classification), however, his results are not sensitive to these types of changes in the empirical design. Thus, our analysis relies only on the ecological diversity indicator.

[28] joshuaproject.net [30 Oct 2014]
The Joshua Project is a research initiative (founded in 1995) that seeks to support Christian missionary activities among the least reached ethnic groups worldwide. They keep an up-to-date freely available database on the size of ethnic and linguistic groups and their religious affinity.

Table 2.3 The expected sign of the key variables in testing Hypotheses 1 to 5

	key variable	name in tables	expected sign
Hypothesis 1	socio-economic development of indigenous societies	socio-economic development	+
Hypothesis 2	share of the groups within the country population	ln(pop share)	+
Hypothesis 3	the age of the Bible translation	bible age	+
Hypothesis 4	the intensity of missionary activity	missions per 100km^2	+
Hypothesis 5	former British colony	British	+

2.4 Empirical results and discussion

2.4.1 Hypothesis testing

The hypotheses discussed in Section 2.2.4 are tested with an ordered logit model (Table 2.4) which dependent variable is the reversed EGIDS scores (Table A2.1) and the main independent variables are those listed in Table 2.3. Each column of Table 2.4 relies on a different version of the socio-economic development variable of indigenous societies (introduced in Section 2.3.3). The specifications contain some additional controls for the spread of Islam and certain geographical and climatic circumstances which have been identified as important sources of ethnolinguistic diversity (Table 2.2).

Table 2.4 supports three out of the five hypotheses. Higher local socio-economic development before the European dominance, higher relative group share within the country population and early Bible translation are indeed found to be positively associated with the current status of languages. The concentration of missionary activities and the nationality of the colonizer do not seem to matter. (However, their indirect role in explaining language status is discussed in the following sub-sections.) There are only two significant geographical control variables. Languages located in areas with lower mean precipitation and further away from the Equator are likely to exhibit higher status. The positive coefficient of the mean temperature is not robust across specifications.

The remaining of the empirical section aims to go beyond these relatively easily identifiable linkages and reveal the factors that influence the dependent variable indirectly, i.e. via a third factor.

Table 2.4 Determinants of language status (ordered logit models with OLS estimation method)

	Specification 1	Specification 2	Specification 3
socio-economic development	**0.342**** **(2.382)**	**0.347**** **(2.228)**	**0.293*** **(1.647)**
ln (pop share)	**0.580***** **(8.247)**	**0.616***** **(7.950)**	**0.654***** **(6.949)**
Bible age	**0.013***** **(3.989)**	**0.015***** **(4.012)**	**0.014***** **(3.260)**
missions per 100km^2	2.642 (0.710)	2.817 (0.717)	2.482 (0.736)
British	0.247 (0.579)	0.322 (0.717)	0.428 (0.748)
French	0.082 (0.182)	-0.154 (-0.321)	-0.424 (-0.666)
Belgian	-0.315 (-0.482)	-0.166 (-0.240)	0.445 (0.503)
Portuguese	-0.049 (-0.074)	-0.050 (-0.070)	0.357 (0.381)
ecological diversity	-0.705 (-1.236)	-0.791 (-1.218)	-1.075 (-1.279)
mean altitude	0.001 (1.414)	0.001 (0.763)	0.001 (0.950)
std of altitude	-0.000 (-0.098)	-0.000 (-0.129)	-0.001 (-0.123)
mean precipitation	**-0.006**** **(-2.487)**	**-0.006**** **(-2.510)**	**-0.006**** **(-2.487)**
std of precipitation	0.024 (1.282)	0.020 (1.087)	0.011 (0.571)
mean temperature	**0.022**** **(2.237)**	0.015 (1.518)	0.007 (0.624)
std of mean temperature	0.007 (0.137)	0.005 (0.094)	-0.010 (-0.141)
ln(distance to equator)	**0.506***** **(2.947)**	**0.506***** **(2.947)**	**0.506***** **(2.947)**
ln(distance to coast)	0.022 (0.147)	0.079 (0.447)	-0.322 (-0.518)
ln(distance to inland water)	-0.086 (-0.942)	-0.009 (-0.088)	-0.182 (-1.416)
islam	-0.286 (-0.566)	-0.644 (-1.079)	-0.744 (-1.100)
number of observations	343	282	195
Pseudo-R^2	0.228	0.256	0.289

Note: Robust t-statistics are in parentheses. *, **, *** label significance at 10%, 5%, and 1% level respectively. Estimated cuts in specification 1 are 1.931, 5.506, 8.731, 9.286, 10.491 and 11.238. Estimated cuts in specification 2 are 0.758, 4.982, 8.331, 8.911, 10.171 and 10.787. Estimated cuts in specification 3 are -3.516, 0.902, 4.772, 5.32, 6.734 and 7.496.

2.4.2 The development of African societies and the contact with Europeans

The development of traditional African societies is measured from the Ethnographic Atlas (Murdock 1967), which intends to reflect the characteristics of societies as they were in the 19th century just before the Scramble for Africa. However, since the

majority of information is obtained from the first half of the 20th century (variable 102 in the Corrected Ethnographic Atlas shown in Figure 2.3), we argue that the Atlas is very likely to encounter precolonial and early colonial European influence to some extent.[29] We use the number of missions on the area of each group per 100 km^2 to proxy the pre-colonial and early colonial European influence. Data are obtained from the Roome's map (1924) which provides a snapshot on the geographical concentration of missionary activities in the early 1920s. The OLS regression model presented in Table 2.5 reinforces our expectations: the number of missions per 100 km^2 is positively associated with the socio-economic development measures (introduced in Section 2.3.3) even after controlling for the decade of the survey and some geographical and climatic variables that are assumed to be important determinants of the location of missions, the early contact with Europeans and the societal organization of traditional societies.

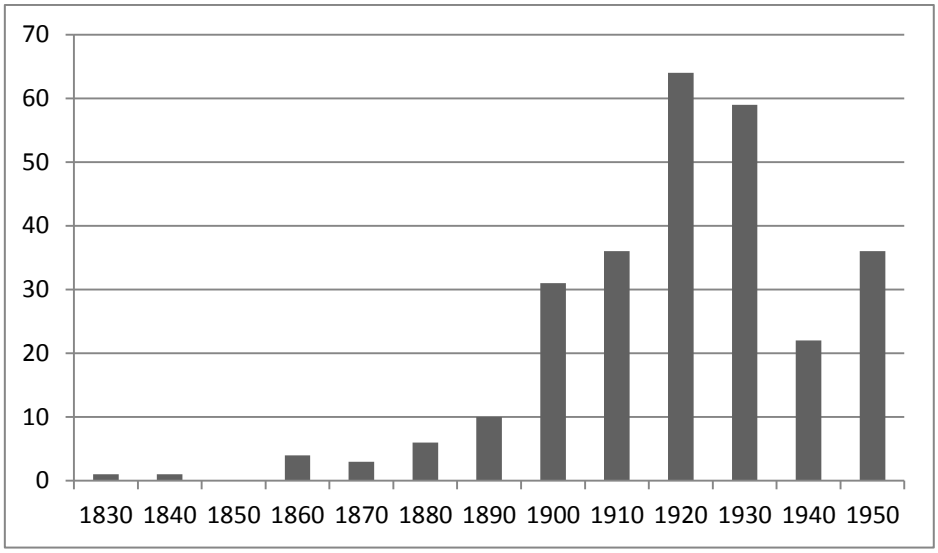
Although until now we have concentrated on the potential effects of European contact on indigenous societies, historical studies argue that the characteristics of local groups also determined the spread of missions (Johnson 1967). If simultaneity is indeed present in the regression of the estimated level of development on the number of missions per 100km^2, the observed positive relationship between the intensity of European contact and the development levels in Table 2.5 is biased. We use instrumental variable technique (two-stage least squares (2SLS)) to identify the causal effect of European contact on the development of indigenous societies (Table 2.6 and 2.6b).

We use the natural logarithm of the distance to the coast and the identity of the colonizer as additional instruments. These variables serve as proper instruments since they are exogenous, and they all affected the probability that missionaries settled in an area, while they are not influenced by the socio-economic development level of an ethnic group. The distance to the coast is assumed to capture the part of European influence that is associated with trading before the colonial era. It has been shown by a number of studies that Europeans came into contact with ethnic groups satiated near the coastal regions first, and these groups were the easiest and most logical destination for early missionaries as well. We also argue that the socio-economic development is associated with the distance from the coast only via trade and contact with Europeans. According to the Ethnographic Atlas, in the majority of societies fishing was not the main economic sector in the 19th century which supports our assumption that the main function of the ocean was to facilitate trade with Europeans. Hence, the distance to coast should be exogenous to the level of development, especially with the

[29] The only study that uses the date of observations provided by Murdock (1967) in a similar way than this chapter does is Henderson and Whatley (2014). By computing the number of years between the date when an ethnic group was colonized and observed, they provide evidence that the duration of colonization contributed to the shift in gender roles, the increasing relative position of men in lineage and inheritance systems, but also contribute to the reduced importance of polygyny in African societies.

proximity to fresh water, the share of Islam as the proxy to trans-Saharan trade and ecological diversity as the proxy for potential reward from trading activities being included in the specifications separately. The colonizer dummies capture the part of the spread of missions that is associated with colonial policies. These variables serve as exogenous instruments since the partition of Africa was mostly a result of the play among European powers, and a region was claimed by a colonizer not because of its relative degree of development but rather by its availability for colonization. Different colonizers lent different support for missionaries as discussed in Section 2.2.2. Table 2.6 (the first stage results of the 2SLS estimates) reveals that the geographical concentration of missionaries was higher near the cost and in former British and Belgian colonies.

Figure 2.3 the number of observed groups in the Ethnographic Atlas per decade from which information is obtained (variable 102)



Note: The figure represents the number of groups without taking partitioning between countries into account. Number of groups is 273.

Table 2.5 The relationship between the early European contacts measured with the intensity of missionary activities and the socio-economic development of African societies (OLS estimates)

	dep. var: socio-economic development1	dep. var: socio-economic development2	dep. var: socio-economic development3
missions per 100km^2	**2.976**** **(2.141)**	**3.325**** **(2.311)**	**3.838**** **(2.585)**
decade of survey	**-0.006*** **(-1.877)**	-0.005 (-1.373)	**-0.008*** **(-1.763)**
ecological diversity	**0.509*** **(1.666)**	**0.785**** **(2.429)**	0.596 (1.554)
mean altitude	0.000 (0.818)	0.000 (0.637)	0.000 (0535)
std of altitude	-0.001 (-0.754)	-0.002 (-1.380)	-0.002 (-0.800)
mean precipitation	0.002 (1.571)	0.001 (1.117)	0.001 (0.339)
std of precipitation	**0.013*** **(1.883)**	**0.015**** **(2.290)**	0.012 (1.608)
mean temperature	-0.004 (-1.020)	-0.004 (-0.934)	-0.007 (-1.440)
std of mean temperature	0.017 (0.683)	0.029 (1.041)	0.011 (0.347)
ln(distance to equator)	**0.145**** **(2.180)**	**0.167***** **(2.563)**	**0.174*** **(1.834)**
ln(distance to inland water)	**-0.108*** **(-1.840)**	**-0.140**** **(-2.242)**	**-0.200***** **(-2.708)**
islam	**0.723***** **(3.010)**	**0.716**** **(2.399)**	**1.009***** **(1.859)**
constant	**11.665*** **(1.739)**	9.877 (1.267)	**15.153*** **(1.724)**
number of observations	236	193	135
R^2	0.229	0.246	0.273

Note: Robust t-statistics are in parentheses. *, **, *** label significance at 10%, 5%, and 1% level respectively.

Table 2.6 The causal effect of European contact measured with the intensity of missions on the development of African societies (first-stage)

	dep. var: missions per 100km^2	dep. var: missions per 100km^2	dep. var: missions per 100km^2
ln(distance to coast)	**-0.019*** (-3.649)**	**-0.021*** (-3.690)**	**-0.025*** (3.600)**
British	**0.022*** (3.054)**	**0.026*** (3.059)**	**0.038*** (3.069)**
French	0.005 (0.668)	0.005 (0.556)	0.005 (1.088)
Belgian	**0.022*** (1.997)**	**0.027*** (2.090)**	**0.042*** (2.409)**
Portuguese	-0.010 (-0.714)	-0.007 (-0.550)	0.013 (0.960)
year of survey	0.000 (0.738)	0.000 (0.922)	0.000 (0.700)
ecological diversity	0.015 (1.394)	0.016 (1.347)	**0.026* (1.705)**
mean altitude	0.000 (0.593)	0.000 (0.628)	0.000 (0.976)
std of altitude	**-0.00009* (-1.838)**	**-0.0001* (-1.891)**	**-0.0001 (-1.259)**
mean precipitation	0.000 (0.908)	0.00 (0.228)	-0.000 (-0.006)
std of precipitation	-0.000 (-0.572)	0.000 (0.163)	-0.000 (-0.339)
mean temperature	0.000 (0.263)	0.000 (0.353)	0.000 (0.696)
std of mean temperature	**0.002** (2.163)**	**0.022** (2.237)**	**0.005* (1.715)**
ln(distance to equator)	**-0.006** (-2.210)**	**-0.005* (-1.728)**	0.001 (0.276)
ln(distance to inland water)	-0.000 (-0.349)	-0.001 (-0.638)	0.001 (0.276)
islam	0.000 (0.033)	0.001 (0.129)	0.006 (0.469)
constant	-0.008 (-0.040)	-0.087 (-0.365)	-0.118 (-0.335)
Cragg-Donald test of weak identification	13.987 (rel. bias<10%)	12.010 (rel. bias<10%)	10.151 (rel. bias<20%)
number of observations	235	192	135
R^2	0.315	0.329	0.395

Note: Robust t-statistics are in parentheses. *, **, *** label significance at 10%, 5%, and 1% level respectively. The H0 of the Cragg-Donald test is that the additional instruments are weak and lead to a bias. (rel. bias<10%) means that the bias of the 2SLS coefficients relative to the OLS coefficients is less than 10%.

Table 2.7 The causal effect of European contact measured with the intensity of missions on the socio-economic development of African societies (IV estimates)

	dep. var: socio-economic development1	dep. var: socio-economic development2	dep. var: socio-economic development3
missions per 100km^2	**11.886*** (3.100)**	**12.248*** (3.176)**	**11.953*** (3.169)**
year of survey	**-0.007* (-1.940)**	-0.006 (-1.434)	**-0.008* (-1.927)**
ecological diversity	0.370 (1.189)	**0.617* (1.855)**	0.273 (0.675)
mean altitude	**0.0002* (1.689)**	0.000 (1.509)	0.000 (1.340)
std of altitude	-0.001 (-0.851)	-0.002 (-1.436)	-0.002 (-0.870)
mean precipitation	0.002 (1.416)	0.001 (1.140)	0.000 (0.173)
std of precipitation	**0.013* (1.932)**	**0.014** (2.186)**	0.013 (1.576)
mean temperature	-0.003 (-0.714)	-0.003 (-0.620)	-0.007 (-1.537)
std of mean temperature	0.012 (0.511)	0.022 (0.825)	0.005 (0.170)
ln(distance to equator)	**0.197*** (2.872)**	**0.212*** (3.123)**	**0.266*** (2.848)**
ln(distance to inland water)	**-0.129** (-2.243)**	**-0.162*** (-2.619)**	**-0.232*** (-3.310)**
islam	**0.878*** (3.674)**	**0.915*** (3.101)**	**1.116*** (3.458)**
constant	**11.822* (1.718)**	10.039 (1.248)	**16.36* (1.816)**
Hansen J-test (df=4, p-value)	0.532	0.679	0.298
number of observations	235	192	135
R^2	0.126	0.128	0.156

Note: Robust t-statistics are in parentheses. *, **, *** label significance at 10%, 5%, and 1% level respectively. The H0 of the Hansen test is that the additional instruments are exogenous and valid.

The second-stage results from the 2SLS specifications (Table 2.7) are similar to those of the OLS estimates in Table 2.5, however, we find a much larger causal effect of the European contact measured with the intensity of missionary activities on socio-economic development than with the OLS method. This may be not only because of removing the effect of simultaneity but also because of removing the effect of measurement errors. Namely, if there are random measurement errors in the explanatory variable (number of missions per 100km^2) then the coefficient estimated by OLS will be biased toward zero. Since our instruments are not correlated with these measurement errors, instrumentation will reduce the bias as well.

The positive significant coefficient of ecological diversity in column 2 is in line with Fenske (2014) who finds that due to potential higher rewards from trade, societies located at the border of ecological zones exhibit higher development (measured with

variable 33 (jurisdictional hierarchy beyond community level) of the Corrected Ethnographic Atlas (Gray 1999).

It is interesting that some geographical factors that are previously documented to determine ethnolinguistic diversity are also found to be associated with socio-economic development. Higher variation in precipitation and distance to the Equator not only supports low diversity (Michalopoulos and Papaioannou 2012, Cashdan 2001) but seem to be positively associated with socio-economic development. This finding is in accordance with the anthropological argumentation proposed by Nettle (1999), that instability in climatic condition encourages cooperation which might be translated to higher societal complexity.

The finding that societies located closer to lakes and rivers (inland water) exhibit higher development suggests that the proximity of fresh water as a scarce resource might have encouraged more complex social organization. Issues related to the management of scarce resources such as water have inspired many scientific fields from economic and social history (Adams and Anderson 1988) through collective action studies (Garrido 2011) to contemporary politics (Peters 1994).

The positive significant coefficient of the current share of Muslims within ethnic groups (a proxy for Arab influence) can be considered as the reinforcement of the special importance of the trans-Saharan trade in the emergence of centralized kingdoms.

Table 2.7 suggests that although the concentration of missionaries does not affect the status of languages directly, it still has indirect positive effect through promoting the socio-economic development level of local groups which turned out to be significant in Table 2.4. The impact of the colonizer dummies can be interpreted similarly. Since the British and Belgian colonial policy fostered the spread of missions to a greater extent compared to independent areas and territories ruled by the French and Portuguese, the nationality of the former colonizer is found to have a persistent indirect effect on language status. Geography and climate seem to play a more important role in determining the intensity of missions and the development of local societies (Table 2.6 and 2.7) than in enhancing language development (Table 2.4).

2.4.3 The share of linguistic groups within the country population

The relative size of linguistic groups is one of the key factors which are assumed to determine the status of languages. Since historical data are not available, this study is based on the Joshua Project's database. However, since it shows the current distribution of linguistic groups, the effects of geography and history are likely to be encountered in the data. Table 2.8 provides evidence that the current share of linguistic groups is dependent on the historical socio-economic development, the Islamic influence and certain geographical factors. And since the socio-economic development of indigenous groups is found to be dependent on the intensity of

missionary activities (shown in the previous section), the share of language groups within current country borders are indirectly influenced by it as well.

Table 2.8 Determinants of population share (OLS)

	Specification 1	Specification 2	Specification 3
socio-economic development	**0.631*** (5.567)**	**0.712*** (5.649)**	**0.643*** (4.051)**
islam	**0.986*** (2.665)**	**0.900** (2.050)**	0.529 (0.979)
British	-0.187 (-0.463)	-0.398 (-0.877)	-0.315 (-0.572)
French	0.112 (0.265)	-0.154 (-0.315)	-0.498 (-0.858)
Belgian	-0.857 (-1.552)	**-1.211** (-1.975)**	**-1.468* (-1.877)**
Portuguese	-0.626 (-0.884)	-0.776 (-1.061)	-1.166 (-0.686)
ecological diversity	-0.415 (-0.863)	-0.190 (-0.348)	-0.071 (-0.102)
mean altitude	**0.001** (1.981)**	**0.001* (1.779)**	0.001 (1.081)
std of altitude	-0.003 (-1.154)	-0.001 (-0.435)	-0.004 (-1.145)
mean precipitation	0.002 (1.056)	0.003 (1.253)	0.002 (0.645)
std of precipitation	**0.029** (2.113)**	**0.024* (1.719)**	**0.031* (1.911)**
mean temperature	**0.013** (1.983)**	**0.014* (1.909)**	0.012 (1.311)
std of mean temperature	0.043 (1.021)	0.019 (0.400)	0.071 (1.104)
ln(distance to equator)	**-0.182* (-1.883)**	**-0.205* (-1.948)**	-0.162 (1.101)
ln(distance to coast)	**-0.413*** (-3.871)**	**-0.402*** (-3.445)**	**-0.329* (-1.968)**
ln(distance to inland water)	-0.081 (-0.832)	-0.082 (-0.728)	-0.173 (-1.454)
constant	**-4.495** (-2.100)**	**-4.558* (-1.836)**	-3.834 (-1.245)
number of observations	343	282	195
R^2	0.236	0.248	0.241

Note: Robust t-statistics are in parentheses. *, **, *** label significance at 10%, 5%, and 1% level respectively.

2.4.4 Counterfactual analyzes

Having identified the factors which have directly or/and indirectly determined the current status of languages, we conduct counterfactual analysis to find out if the language status pattern in Sub-Saharan Africa would be different if there was no European influence. This technique is designed to overcome the problem that we cannot carry out historical experiments. Counterfactual analysis has been used in several economic studies to measure the effects of historical events. Nunn and Qian

(2011) reveal that the introduction of the potato is responsible for about one-quarter of the population growth and urbanization in the Old World between 1700 and 1900. Fernihough and O'Rourke (2014) show that the introduction of coal-using technologies explains about 60% of the increase in the European city population between 1750 and 1900.

Our goal is similar to these works. We aim to compare two situations: the pattern of the current status of African languages with and without European influence. Using the results presented in Tables 2.4 to 2.8, we can reconstruct the language situation without missionary activities, Bible translation and colonization by following the steps detailed below.

The socio-economic development of indigenous societies measured from the Ethnographic Atlas and the share of language groups are positively related to the current status of languages (Table 2.4). However, the intensity of missionary activities affected the development of indigenous groups (Table 2.7) which has contributed to higher group share within the country (Table 2.8). Moreover, the nationality of the colonizer is found to exert a significant influence on the relative size of language groups in certain cases (Table 2.8). In order to see how the status of languages would be without the European influence, these direct and indirect effects should be filtered out.

The indigenous socio-economic development and the ln(pop share) variables without European influence are computed as shown in Eq. 2.1 and Eq. 2.2, respectively. Coefficients are taken from Table 2.7 and Table 2.8. The formulas are presented in the case of the first type indigenous socio-economic development variable (specification 1 of Table 2.1) introduced in Section 2.3.3.

Let us consider the case of the Ambo in Angola (former Portuguese colony) where the number of missions per 100 km^2 is 0.007. The value of the first type socio-economic development estimated from the Ethnographic Atlas is 0.919 which reduces to 0.837 (0.919-11.886*0.007) if we filter the effect of missions out (Eq. 2.1). In other words, the presence of missions contributed to 0.082 (0.919-0.837) higher early socio-economic development in the case of the Ambo group. According to Eq. 2.2, the natural logarithm of the Ambo group share within Angola would be -3.381 (-3.956-0.631*0.082+0.626) instead of -3.956 if there was no European influence. The distribution of the socio-economic development with and without European influence is presented in Figures A2.3a to A2.3c. The order of the ten most developed indigenous societies with and without European influence is shown in Table A2.6.

$$socioeconomic\_development1(without) = socioeconomic\_development1 - 11.886 * missions/100km^2 \quad \text{(Eq. 2.1)}$$

$$\ln(pop\_share)without = \ln(pop\_share) - 0.631*(socioeconomic\_development1 - socioeconomic\_development1(without)) + 0.187*British - 0.112*French + 0.857*Belgian + 0.626*Portuguese \quad \text{(Eq. 2.2)}$$

The language status without European influence is computed according to Eq. 2.3. Coefficients are taken from Table 2.4. The formula is shown only in the case of the first type socio-economic development variable. Since language status is a discrete dependent variable which is estimated with an ordered logit model, the computation of the language status without European influence is not straightforward. Instead of the observed EGIDS, our starting point is the predicted status which is a continuous variable. Instead of removing the effect of Bible translation (bible age) completely, we consider the difference between the actual and the average bible age.

$$\begin{aligned} &language\_status(without) = \\ &language\_status(predicted) - 0.342 * (socioeconomic\_development1 \\ &-socioeconomic\_development1(without)) \\ &-0.580 * (\ln(pop\_share) - \ln(pop\_share)without) \\ &-0.013 * (bible\_age - 73) - 0.247 * British - 0.082 * French \\ &+0.315 * Belgian + 0.049 * Portuguese \end{aligned} \qquad \text{(Eq. 2.3)}$$

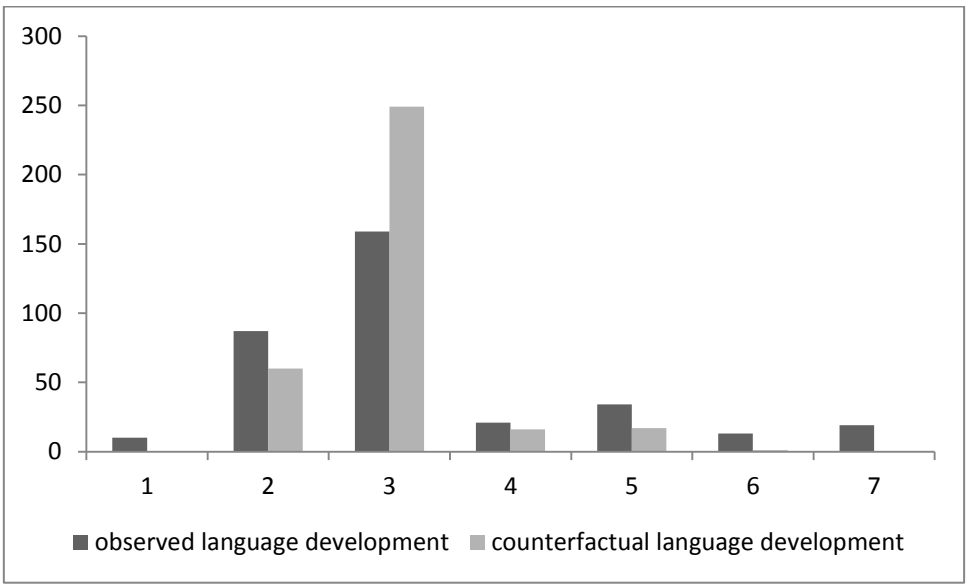
Again, we take Ambo in Angola to illustrate the use of Eq. 2.3 in practice. Ambo is a language of wider communication (reversed EGIDS is 5) which has had a Bible translation since 1878. The predicted (continuous) language status is 9.069. Without European influence, this value would be 8.558 (9.069-0.342*0.082-0.580*(-0.585)-0.013*(136-73)+0.049) which equals 3 on the reversed EGIDS according to the estimated cut values listed under Table 2.4. This result can be interpreted that Ambo in Angola would have a lower status today without European influence. Or, in other words, Ambo has benefitted from missionary activities and colonization.

The effect of early European contact and colonialism on the distribution of language status in the case of the first socio-economic development variable is presented in Figure 2.4. The same figures for the second and third type socio-economic variable are found in Table A2.4a and A2.4b. According to our models, the distribution of the status of languages would be similar to the current distribution in terms of the mode: category 3 which is EGIDS 5 (developing) in the original coding is the most populous. However, none of the languages belong to category 7 (EGIDS 1 (national)) and category 1 (combined group from EGIDS 6b, 7 and 8a (threatened)) anymore when socio-economic development is measured with the first two types. The only language that falls into the highest development category 7 when socio-economic development of traditional societies is measured with the third type variable is the Sotho (Lesotho and South Africa). Thus, Figure 2.4 (and Figures A2.4a and A24.b) suggests that the colonial rule and missionary activities altogether increased polarization in the distribution of the status of languages in Sub-Saharan Africa which can be traced in the smaller kurtosis and a relative heavy-tailedness of the distribution when European influence is not filtered out.

In order to indicate how much European influence affected the distribution of the status of languages, we apply the 1-Herfindahl index[30], which is often used in ethnolinguistic diversity measurement. In our case, 1-Herfindahl index can be interpreted as the probability that two randomly selected languages in our dataset have different status. The value of the 1-Herfindahl index related to the distribution of the current status of languages in Figure 2.4 is 0.702 (dark grey columns), and 0.515 in the case of the counterfactual situation (light grey columns). The difference between the two values suggests that the European influence increased the probability that two languages are of different status by about 20 percentage points. (The computed values related to Figure A2.4a and A2.4b are similar.)

Since language status determine the potential of its speakers for participating in the labor market and political decision making, our results are broadly related to the issue of individual well-being, economic development and inequality. What we have found can be interpreted that early contact with Europeans, the selective Bible translation strategies of missions and colonial policy contributed to higher inequality across groups through influencing their language development paths. The counterfactual analysis also suggests that if the current status of languages was dependent only on their traditional socio-economic characteristics, the gaps between the least and the most linguistically developed groups would be smaller today.

Figure 2.4 The distribution of the observed and the counterfactual language status



[30] The Herfindahl-index (Herfindahl 1950) is computed as $\sum_{j=1}^{7} g_j^2$ where g_j is the share of languages within the sample with status j.

2.5. Conclusion

This study aims to reveal the historical determinants of the current status of languages in Sub-Saharan Africa. The development of languages is measured with the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Lewis and Simons (2010)) from Ethnologue (Lewis et al. 2014) which evaluates the vitality of languages according to five key concepts. The sources include the map by Murdock (1959) displaying the geographical location of indigenous societies in the 19th century, the Ethnographic Atlas (Murdock 1967) on the socio-economic characteristics of ethnic groups, the Roome map (1924) that contains the location of early missions, different sources on the year of the first Bible translation, the Joshua Project on the share of ethnic groups within country borders, the nationality of colonizer before independence, and the share of Islam within each group. The empirical models also control for geographical and climatic factors (distance to equator, coast and inland water, the mean and standard deviation of temperature, altitude and precipitation, and ecological diversity).

Although the Ethnographic Atlas is a widely utilized anthropological source in development studies, the way we apply this source is different from the traditional approaches in two aspects. First, most economic studies that show the persistence and the long term impact of past social traits on current development focus on variable 33 'jurisdictional hierarchy beyond the community level'. Using a general structural equation modeling technique, this chapter utilizes several variables related to economic and social organization from the Ethnographic Atlas to estimate the (assumed) underlying latent socio-economic development of indigenous societies, which is, unlike the original ordered variables, measured on a continuous scale. We estimate three latent socio-economic development variables with three different observed variables sets which, due to the high share of missing values in the Ethnographic Atlas, results in observation numbers of 343, 282, and 195. Second, we also argue that the Ethnographic Atlas does not purely provides information on the 'aboriginal' traits of African societies, but incorporates the effects of contacts with Europeans (earlier via trade and missions and later through colonial practices and policies) to some extent. Using an instrumental variable technique (two-stage least squares) we separate the effect of European influence measured with the intensity of missionary activities on the development of indigenous societies (estimated with the latent variable model).

When analyzing the main determinants of the current status of languages, we find that the socio-economic development of traditional societies, the age of Bible translation, the share of ethnic groups within the country are positively associated with the dependent variable. However, the number of missions per 100 km^2 and the colonizer dummies do not yield significant coefficients. Since the colonizer dummy is proved to be an important determinant of the intensity of missionary activity, we

argue that colonial history has indirect effect on language status through determining the spread of Christian missions.

Comparing the distribution of the current status of languages with the counterfactual distribution if no European influence had taken place suggests that missionary activities and colonialism have a persistent linguistic effect that contributed to higher polarization in the language status distribution, thus higher socio-economic inequality among linguistic groups in Sub-Saharan Africa.

Appendix 2A

Table 2A.1

The description of the dependent variable and the distribution of sample languages per status

original EGIDS scale	status	explanation	number of languages (total: 389)	percent	number of African languages in Ethnologue (total: 2735)	percent	recoded value for the empirical analysis (reversed EGIDS)
0	international	The language is widely used between nations in trade, knowledge exchange, and international policy.	0	0	0	0	-
1	national	The language is used in education, work, mass media, and government at the national level.	19	4.88	72	2.63	7
2	provincial	The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.	14	3.6	17	0.6	6
3	wider communication	The language is used in work and mass media without official status to transcend language difference across a region.	36	9.25	109	3.99	5
4	educational	The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.	26	6.68	89	3.25	4
5	developing (written)	The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.	182	46.79	756	27.64	3
6a	vigorous	The language is used for face-to-face communication by all generations and the situation is sustainable.	99	25.45	1223	44.72	2
6b	threatened	The language is used for face-to-face communication within all generations, but it is losing users.	10	3.34 (share of the	199	7.28	1

7	shifting	The child-bearing generation can use the language among themselves, but it is not being transmitted to children.	2	combined category)	86	3.14	1
8a	moribund	The only remaining active users of the language are members of the grandparent generation and older.	1		58	2.12	1
8b	nearly extinct	The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language.	0	0	66	2.41	-
9	dormant	The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency.	0	0	24	0.88	-
10	extinct	The language is no longer used and no one retains a sense of ethnic identity associated with the language.	0	0	36	1.32	-

Note: since our analysis is concerned with explaining language status instead of language endangerment we reversed the scale so that higher values refer to higher status. Thus, the interpretation of the coefficients in the empirical analysis is more straightforward.

Table 2A.2

Variables used for the estimation of the socio-economic development of indigenous societies

variable name	variable number in the Corrected Ethnographic Atlas	#	original coding	recoding	remark
intensity of agriculture	var 28	389	no agriculture (1), casual agriculture (2), extensive or shifting agriculture (3), intensive agriculture, using fertilization, crop rotation, or other techniques to shorten or eliminate fallow period (5), intensive irrigated agriculture (6)	1 and 2 -> 0 3 -> 1 5 and 6 -> 2	Our sample does not include category 4 (horticulture, vegetal gardens or groves, fruit trees) of v28. The aim of the recoding strategy is to distinguish between extensive and intensive agriculture, assuming that intensive agriculture indicates higher social and economic development.
jurisdictional hierarchy beyond community level	var 32 (only the second digit)	386	no levels (1), one level (2), two levels (3), there levels (4), four levels (5)	1 -> 0 2 -> 1 3 -> 2 4 and 5 -> 3	Category 4 and 5 are collapsed since there is only one observation in category 5.
class stratification	var 67	354	absence among freemen (1), wealth distinctions (2), elite (based on control of land and other resources) (3), dual (hereditary aristocracy) (4), complex (social classes) (5)	1 -> 0 2 -> 1 3 -> 2 4 -> 3 5 -> 4	
succession of the office of local headmen	var 73	319	patrilineal heir (1), matrilineal heir (2), appointment by higher authority (non-hereditary) (3), seniority or age (non-hereditary) (4), influence, wealth or social status (non-hereditary) (5), election or other formal consensus (non-hereditary) (6), informal consensus (non-hereditary) (7), absence of any such office (9)	9 -> 0 1 and 2 -> 1 3, 4, 5, 6, 7 -> 2	The recoding strategy is aimed to distinguish between hereditary and non-hereditary systems. Societies with no such office are considered as the reference category.
high god	v34	263	absent or not reported (1), not active in human affairs (2), active in human affairs but not supportive in human morality (3), supportive of human morality (4)	1->0 2->1 3-> 2 4->3	

Table 2A.3

Correlation coefficients between the different estimated socio-economic development variables

	dev 1	dev 2	dev 3
dev 1	1		
dev 2	0.990 (194)	1	
dev3	0.983 (136)	0.993 (136)	1

Note: the number of observations is reported in parentheses. All coefficients are significant at 1%.

Table 2A.4

Spearman rank correlation coefficients between the estimated latent socio-economic development variables and their components

	dev 1	dev 2	dev 3
agricultural intensity	0.301	0.332	0.390
jurisdictional hierarchy (beyond community level)	0.904	0.847	0.850
class stratification	0.863	0.908	0.915
authority succession	-	-	0.225
high god	-	-	0.324
Number of obs	237	194	136

Note: All coefficients are significant at 1%.

Table 2A.5

The ten most developed indigenous societies with and without European influence according to the three indigenous socio-economic development measures

order	socio-economic development1	socio-economic development1 without European influence	socio-economic development2	socio-economic development2 without European influence	socio-economic development3	socio-economic development3 without European influence
1	Rundi	Amhara	Amhara	Amhara	Amhara	Amhara
2	Ruanda	Songhai	Rundi	Songhai	Songhai	Songhai
3	Amhara	Rundi	Ruanda	Rundi	Rundi	Rundi
4	Songhai	Oyo Yoruba	Songhai	Oyo Yoruba	Ruanda	Tigrinya
5	Oyo Yoruba	Ruanda	Oyo Yoruba	Tigrinya	Tigrinya	Kanuri
6	Hunde	Kanuri	Tigrinya	Kanuri	Oyo Yoruba	Oyo Yoruba
7	Tigrinya	Fur	Nupe	Ruanda	Nupe	Fur
8	Lozi	Janjero (Yemsa)	Ganda	Fur	Kanuri	Janjero (Yemsa)
9	Sotho	Kafa	Kanuri	Janjero (Yemsa)	Wolof	Ruanda
10	Nyoro	Sotho	Merina (dialect of Malagasy Plateau)	Sotho	Fur	Nupe

Table 2A.6

Variable description

variable name	description	variable type	source
language status	The status of each language as understood by the EGIDS (Lewis and Simons 2010). The original and recoded values are presented in Table A.1. For more information consult http://www.ethnologue.com/about/language-status	categorical (ordered)	Ethnologue (Lewis et al. 2014)
British, French, Belgian, Portuguese	It labels ethnic groups that were assigned to countries ruled by the British, French, Belgian and Portuguese. If a group was partitioned between countries, it is labeled in each country accordingly.	dummy/binary	Bertocchi and Canova (2002)
socio-economic development of indigenous societies	Estimated with GSEM techniques from variables in Table 2.A as shown in Table A.4 described in Section 3.	continuous	Ethnographic Atlas (Murdock 1967)
ln(pop share)	The share of each ethnic group within country population.	continuous between 0 and 1	Joshua Project
bible age	The number of years between the year that the Bible was translated to a certain language and 2014 (present).	continuous with integer values	Groves (1964), Ethnologue (2014), worldbibles.com
missions per 100km^2	The number of missionary locations on the territory of each ethnic group presented in Murdock (1959) map.	continuous	Roome (1924), Nunn (2008 and 2010)

islam	The share of each ethnic group affiliated with the Islam religion.	continuous between 0 and 1	Joshua Project
ln(distance to equator)	The distance of the center of each ethnic group to the Equator (in km).	continuous	The shapefile (ESRI data format) containing the Equator is downloaded from the Natural Earth website. http://www.naturalearthdata.com/features/
ln(distance to coast)	The distance of the center of each ethnic group to the coast (in km).	continuous	The shapefile of African coastline is downloaded from the website of the African Marine Atlas. http://omap.africanmarineatlas.org/BASE/pages/coastline.htm
ln(distance to inland water)	The distance of the center of each ethnic group to the closest river or lake (in km).	continuous	Shapefile with rivers is downloaded from http://www.arcgis.com/home/item.html?id=fedf8e234b614ecaac65893f807344f5 The shapefile with lakes is downloaded from
mean temperature	The mean of monthly temperature on the area of each ethnic group between 1950 and 2000 (in 0.1 Celsius).	continuous	Raster data (ESRI grids format, 10 arc-minutes resolution) are from the WorldClim website. For the description of applied methods see http://www.worldclim.org/methods and Hijmans, R. J. et al. (2005)
std of temperature	The standard deviation of monthly temperature within the territory of each group (in 0.1 Celsius).		
mean precipitation	Mean monthly precipitation on the territory of each ethnic group (in mm).		
std of precipitation	The standard deviation of mean precipitation within the territory of each ethnic group (in mm).		
mean altitude	The mean elevation of each ethnic group above sea level (in m).		
std of altitude	The standard deviation of elevation within the territory of each ethnic group (in m).		
ecological diversity	This variable show the probability that two geographical points selected at random within the territory of each ethnic group belong to different ecological zones defined in White (1983).	continuous between 0 and 1	The vegetation map is published by White (1983). The shapefile is available at James Fenske's website https://sites.google.com/site/jamesfenske/data

Note: original and recoded values are understood and presented in the case of categorical variables. Original values column shows only values that are present in our sample. For information on other values consult the Corrected Ethnographic Atlas (Gray 1999).

Figure 2A.1
The location of indigenous ethnic groups (Murdock 1959)

Figure 2A.2
The location of early missions (Roome 1924)

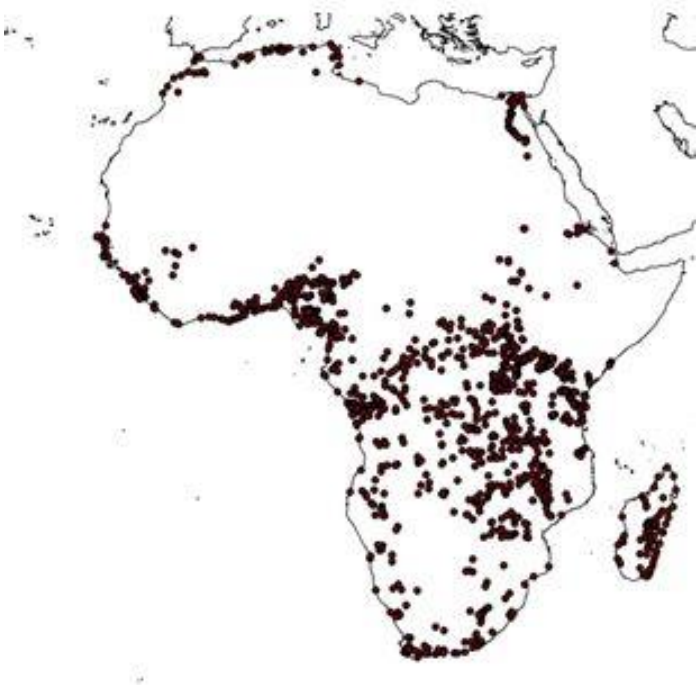


Figure 2A.3a

Kernel density estimates of the distribution of socio-economic development with and without European influence (development variable: socio-economic development 1)

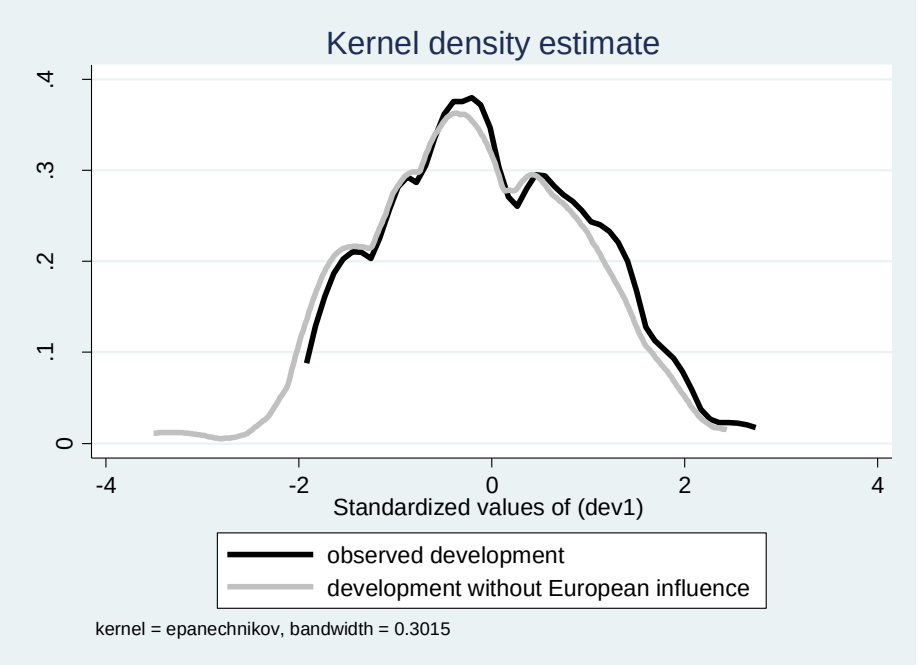


Figure 2A.3b

Kernel density estimates of the distribution of socio-economic development with and without European influence (development variable: socio-economic development 2)

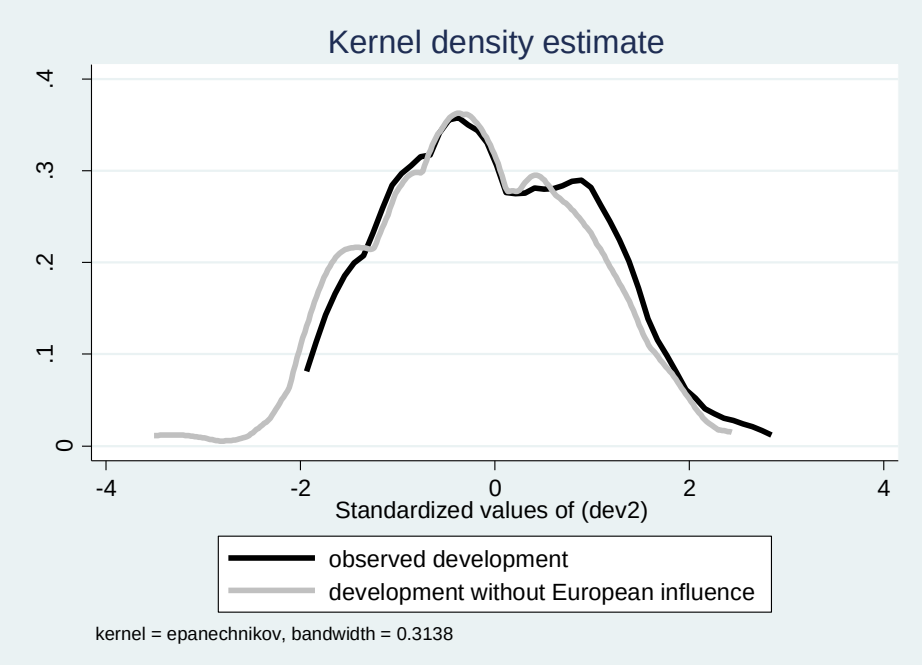


Figure 2A.3c

Kernel density estimates of the distribution of socio-economic development with and without European influence (development variable: socio-economic development 3)

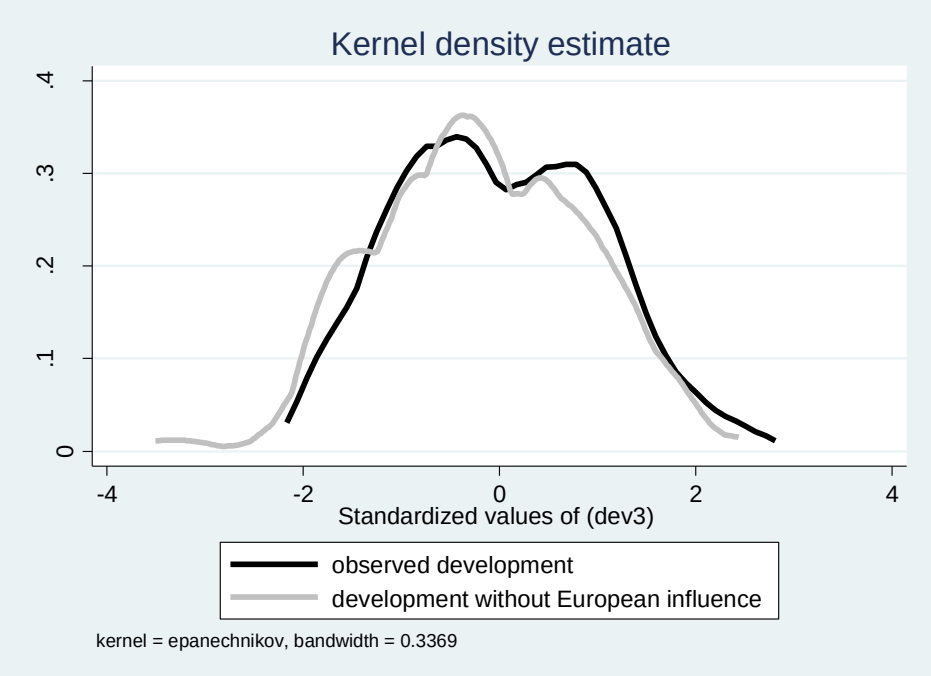


Figure 2A.4a

The distribution of the observed and counterfactual language status (socio-economic development 2)

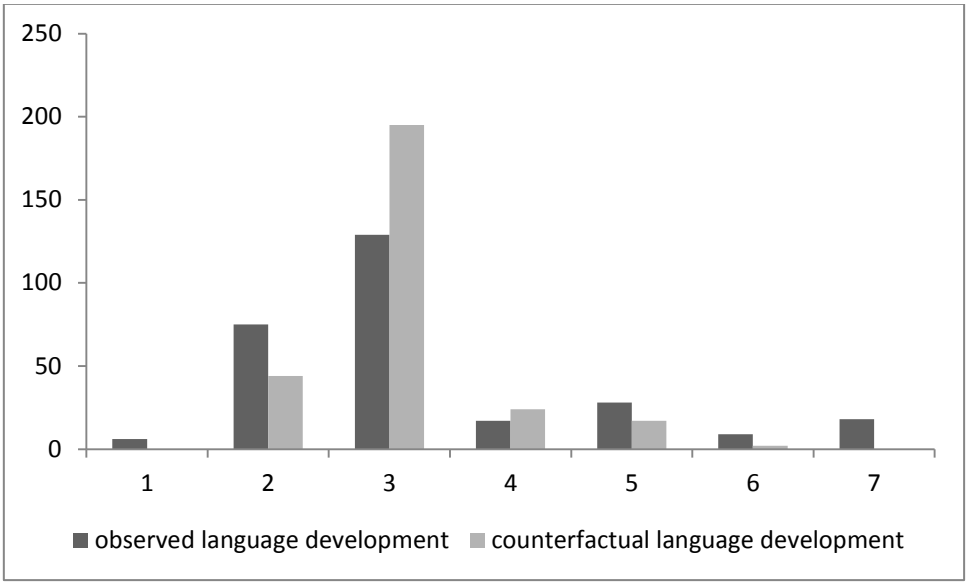
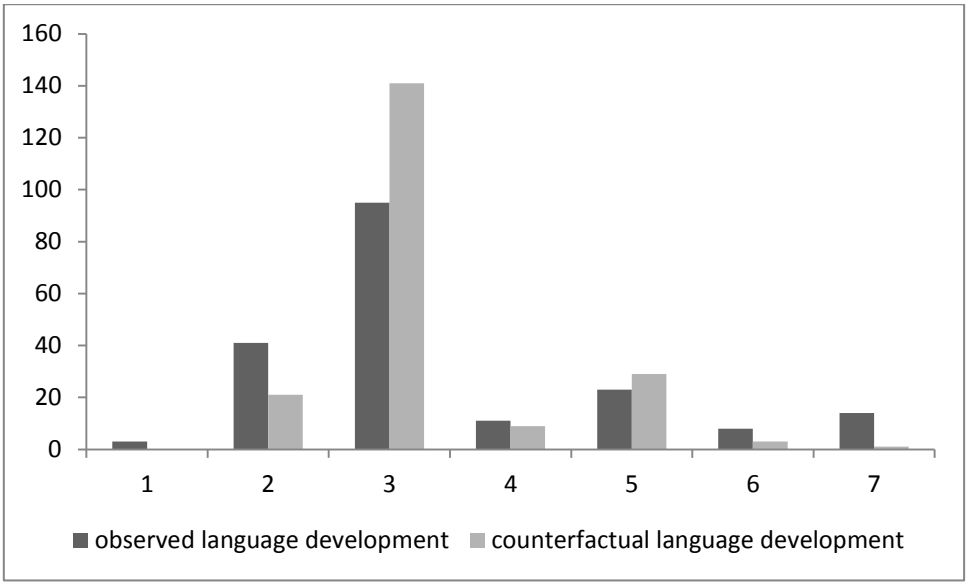


Figure 2A.4b

The distribution of the observed and counterfactual language status (socio-economic development 3)



3 Linguistic situation in twenty Sub-Saharan African countries: A survey-based approach[31]

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 utilizes the 4th round 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, we apply the Index of Communication Potential (ICP) which, according to the author's knowledge, is the first linguistic measure that takes multilingualism into account. And finally, we show how a simple graphic representation of the ICP can be used to visualize 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 colonizers' languages. Our findings are expected to be of interest of scholars engaged in language policy and planning and language-related development issues.

Keywords: Sub-Saharan Africa, linguistic situation, multilingualism, linguistic diversity, linguistic data, Index of Communication Potential, Afrobarometer Survey

[31] This study is forthcoming in African Studies (Taylor and Francis). The author is grateful to Maarten Mous (Leiden University, The Netherlands), Peter Foldvari (Utrecht University, The Netherlands), the two anonymous referees at African Studies, and the participants of the Second Lisbon Meeting on Institutions and Political Economy (2013, Lisbon, Portugal) for their invaluable comments on the Index of Communication Potential and the previous versions of this chapter.

3.1 Introduction

It is well established that Africa is characterized by high ethnic and linguistic heterogeneity (Lewis et al. 2014, Alesina et al. 2003, Akademija 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 and 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) usually provide information on ethnicity, mother tongue or home language, obtaining data on additional languages, which is an essential requirement for analyzing the patterns of multilingualism and language dynamics, is more difficult.

This study attempts to fill this lack within the literature to a certain extent by utilizing the fourth round of the Afrobarometer Survey (2008 and 2009) that contains not only the ethnicity and home language but also the additional languages of more than 27000 individuals in 20 Sub-Saharan African countries. Although the review of the size of ethnic and linguistic 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, we aim to present our findings in a more insightful way. This study predominantly relies on the Index of Communication Potential (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 Index can be applied to visualize 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. Our results are expected to be of interest of language and education planners, economic and political scientists focusing on the development consequences of diversity and multilingualism, and other researchers whose work include designing and interpreting surveys including questions on language or ethnicity.

The chapter is structured as follows. The next section overviews those research fields in which information on the number of speakers is essential. Section 3 discusses the benefits and limitations of the Afrobarometer Survey as a linguistic data source and its comparison with other available materials. Section 4 introduces the Index of Communication Potential (ICP). Section 5 is devoted to a simple graphical representation of the ICP to show the patterns of language use in the 20 sample

countries in a comparative way. Section 6 concludes and discusses how our findings relate to other disciplines.

3.2 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.[32] However, since they are more relevant from the aspects of our study, below we 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 and 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.[33] Since they are proper sources to estimate linguistic diversity (Desmet et al. (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 and Levine 1997, Pool 1972), social capital (Putnam 2007, Alesina and La Ferrara 2002) and the quality of government (Mauro 1995) are negatively, while the probability of internal conflicts (Montalvo and Reynal-Querol 2010 and 2005) is positively associated with ethnolinguistic 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 1968-71 financed by the Ford Foundation and sponsored by local universities was the first large-scale sociolinguistic research initiative in Africa (see Polomé 1982 for a detailed overview). 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

[32] URL of the WALS Project: http://wals.info/, URL of OLAC: http://www.language-archives.org/ [23 March 2015]
[33] http://joshuaproject.net/ [23 March 2015]

in the supplementary material). The 'Language and dialect atlas of Kenya' edited by Heine and 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 supplementary material and in van der Merwe and van der Merwe 2006 pp. 1-2.), which utilize 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 (e. g. religion, age structure, education, segregation index) of their speakers. An additional language project was initiated by the 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 recognized officially. 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 level (detailed overview of the challenges and the results of the project is provided in Muzale and 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 Current Issues of Language Planning and Journal of Multilingual and Multicultural Development collected in Baldauf and Kaplan (2004) and Kaplan and Baldauf (2007). The role of indigenous languages in education is one of the most debated issues (e.g. Rabenoro 2013, Capo et al. 2009). Some works address specific language policy questions such as the violation of language rights (Namyalo and Nakayiza 2014).

And finally, collecting data on language use behavior 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 pp. 93-94). 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 pp. 248-251, Fishman 1991). Hence, the loss of speakers is recognized as a sign of increased endangerment (Lewis and Simons 2010, UNESCO 2003, Fishman 1991). Having acknowledged the social problems associated with language death and linguistic diversity loss, numerous programs have been initiated to identify threatened languages (UNESCO Atlas of the world languages in danger (Moseley 2010), Catalogue of endangered languages under the direction of University of Hawai'i Mānoa and LINGUIST List/Eastern Michigan University) and to reverse the process of language decline across the world (e.g. language development works at SIL International, projects within the UNESCO

Endangered Languages Programme and projects financed by the US National Science Foundation).[34]

3.3 The Afrobarometer Survey as a linguistic data source

3.3.1 The survey

The Afrobarometer Survey (hereafter AB) is an independent, non-political research initiative to map the social, political, and economic atmosphere in Africa. Since it is conducted regularly[35] and provides a representative sample of citizens of voting age[36] and a standard set of questions, the Afrobarometer has recently become an acknowledged source of development-related research (Eifert et al. 2010, Nunn 2010). Unfortunately, additional languages are included only in Round 4; thus, this chapter is limited to the twenty countries[37] and the two consecutive years (2008 and 2009) covered in that wave. In this chapter, 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.

3.3.2 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 second (for instance the Ethnologue (Lewis et al. 2014)), individual-level data allow us to capture a country's typical language repertoire, 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

[34] URL of the SIL International: http://www.sil.org/. URL for the UNESCO Endangered Languages Project: http://www.unesco.org/new/en/culture/themes/endangered-languages/; URL for information on the Catalogue of endangered Languages: http://www.endangeredlanguages.com/; URL for the Moro Language Project financed by the National Science Foundation: http://moro.ucsd.edu/ [23 March 2015]

[35] 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 that covers 36 countries including those in Northern Africa is being processed and digitalised at the moment. Round 6 is under preparation.

[36] 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 into the sample.

[37] Benin, Botswana, Burkina Faso, Cape Verde, Ghana, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mozambique, Namibia, Nigeria, Senegal, South Africa, Tanzania, Uganda, Zambia, and Zimbabwe

(country, region, urban-rural distinction etc.), thus still can be applied for studies with 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.

And finally, 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 and 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.

3.3.3 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 socio-linguistics, 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 analyzed.

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. 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 and Selinker 2008 p. 7., Chuo Kikuu cha Dar es Salaam 2009 p. xii). Although the two concepts are often considered as synonyms in everyday use, due to migration, interethnic 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 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 and Langer 2010, Burton et al. 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 mobilize 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 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 (e.g. 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 that '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, p. 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 Beardsmore 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 et al. 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 we refrain from labelling them 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 pp. 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 (e.g. the distinction between second and foreign languages (Gass and Selinker 2008 p. 7.), or the distinction between second, third and additional languages (Ortega 2009 pp. 5-7.)), which we do not aim to address. 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.

3.3.4 Comparison with alternative sources

In order to overcome the aforementioned shortcomings, our findings are cross-checked against the following alternative sources: Ethnologue (Lewis et al. 2014), the latest available national censuses, literacy reports, Demographic and Health Surveys, the documents of the Organisation Internationale de la Francophonie (OIF), Albaugh (2014), and other available documents on individual countries. Data are presented and discussed in the supplementary material.

The general conclusion that can be derived from the 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 colonizer'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 discussed above, 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 and 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 census 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 census do not include language- or ethnicity-related questions at all.

An additional source of discrepancies 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 much 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 (footnote 2 under the table of Benin in the supplementary material), the reported shares become more comparable across sources.

And finally, we discuss the causes of the high differences across sources related to the use of additional languages. Since they are usually not surveyed directly, we rely on various materials such as the website of the Organisation Internationale de la Francophonie (2010), individual estimates collected in Albaugh (2014), and literacy data from the latest population censuses, literacy reports and the Demographic and Health Surveys 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 (e. g. 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 Demographic and Health Surveys (DHSs) 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 DHSs 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 conduct separate country-wide literacy surveys regularly. But, while Botswana measures reading, writing and oral language skills apart, the Nigerian survey is based on self-report.

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

The strategy of utilizing literacy data to gain more insight in 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 it is demonstrated in the 2003 Botswana Literacy Survey (Central Statistics Office 2005), measured writing, reading and communication skills can differ significantly. Whereas 38% and 34.2% 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% (Central Statistics Office 2005, Table 38, p. 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 our data is difficult, we find that the shares of people with communication and literacy abilities in the former colonizers' languages are quite similar across sources in about the 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 characterized by relatively moderate differences, we find striking anomalies.

3.4 The Index of Communication Potential (ICP)

The need for a linguistic diversity (also called fragmentation and heterogeneity) measure that accounts for multilingualism has long been recognized in linguistics and political science. In an early work that systemizes the possible approaches, 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 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. Buzasi (2015) finds evidence that African people living in regions with higher average Index of Communication Potential, the main measure in this work, are more likely to trust unknown people. Aspachs-Bracons et al. (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 us with 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. We apply the Index of Communication Potential (Buzasi 2015, hereafter ICP)[38] 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 are provided in Appendix 3A. The individual ICP scores (Eq. 3A.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 (Eq. 3A.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

[38] Our index is different from the Q-value of communication potential introduced by Abram de Swaan. Although both indicators attempt to measure the value of language repertoires in terms of the share of a population that can be reached through them, 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 (De Swaan 1993). Later, the Q-value of certain languages and repertoires was computed for Congo/Zaire (De Swaan 1996), Senegal and South Africa (De Swaan 2001).

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 1 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 3B), which are utilized in development studies, are presented in parallel to the ICP in Table 3.1. Using Q79 on ethnic affiliation and Q3 on home languages from Afrobarometer, we compute 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, we do not compute the ethnic diversity measure for this country.

Table 3.1 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 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% of the Malinke people, by 49.7% of the Peulh/Fulfulde group, by 28.3% of the Senoufo/Mianka ethnic group, and by 25.7% of the Soninke/Sarakolle people. The case of Wolof is similar in Senegal: 35.5% of the Serer, 22.3% of the Pulaar/Toucouleur and 19.7% of the Mandinka/Bambara reported Wolof as the primary language at home.[39]

[39] Groups are named and spelled as in the codebooks of the 4th round of the Afrobarometer Survey.

Table 3.1 Ethnic and linguistic fragmentation and the Index of Communication Potential in the Afrobarometer countries

country (the number of respondents)	ethnic fragmentation	linguistic fragmentation	ICP (individual standard deviation)
Benin (1200)	0.825	0.816	0.581
Botswana (1200)	0.923	0.407	0.984
Burkina Faso (1200)	0.688	0.703	0.602
Cape Verde (1264)	-	0.005	0.995
Ghana (1200)	0.755	0.718	0.751
Kenya (1104)	0.890	0.892	0.917
Lesotho (1200)	0.888	0.040	1.000
Liberia (1200)	0.888	0.885	0.598
Madagascar (1350)	0.826	0.020	1.000
Malawi (1200)	0.781	0.728	0.884
Mali (1232)	0.839	0.719	0.803
Mozambique (1200)	0.874	0.872	0.697
Namibia (1200)	0.705	0.701	0.816
Nigeria (2324)	0.856	0.876	0.622
Senegal (1200)	0.701	0.605	0.892
South Africa (2400)	0.866	0.855	0.606
Tanzania (1208)	0.954	0.950	0.991
Uganda (2431)	0.896	0.896	0.484
Zambia (1200)	0.884	0.872	0.663
Zimbabwe (1200)	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. The number of respondents is presented in parentheses.

The second main conclusion derived from Table 3.1 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. The 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% 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% of the Ndebele speakers who represent more than 10% of the population tends to be monolingual and 43.47% 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 eleven official languages in South Africa.

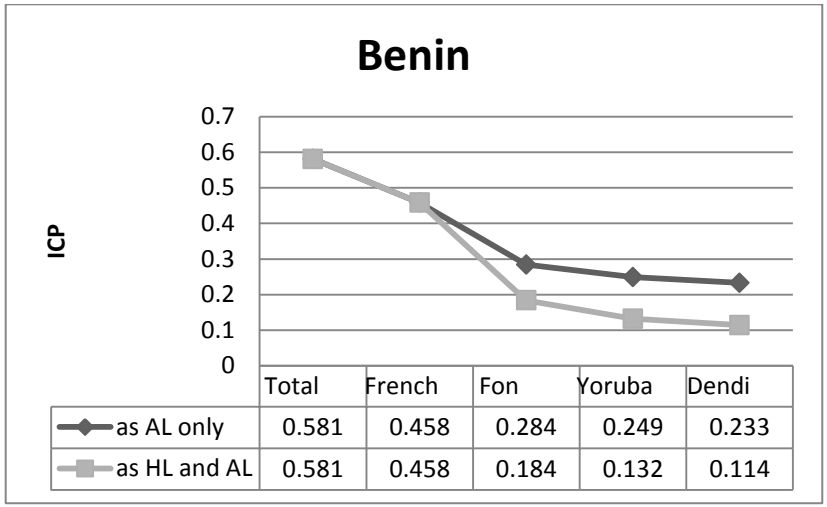
3.5 A graphic representation of the ICP

In the last part of this study we show 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, we sort languages by their size as an additional language in each country and recalculate 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 grey lines in Figure 3.1 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 grey 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 the listed languages were not spoken as an additional language (dark grey) or were not spoken at all (light grey). The magnitude of the decrease and the difference between the dark and light grey 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 3.1.

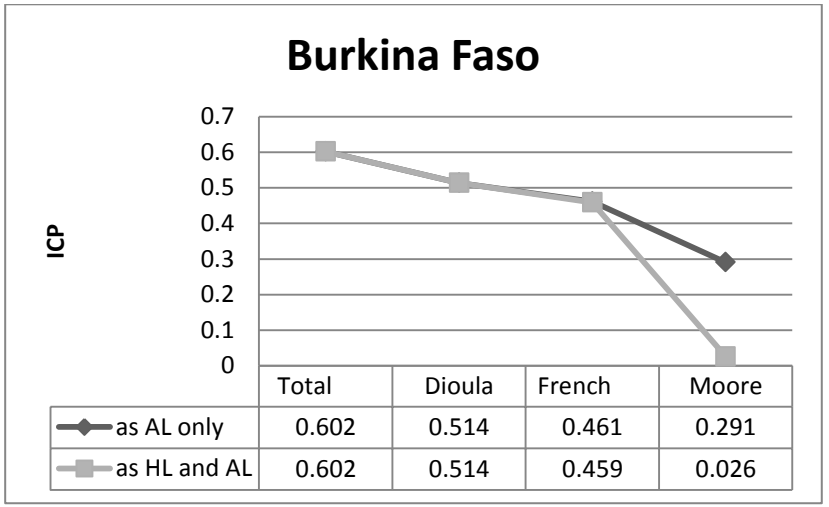
For instance, let us consider the case of Ghana in Figure 1 and in the supplementary material. The order of languages according to their reported frequency as additional language is English (47.39%), Akan (31.57%), Ga/Dangme (11.04%), and Ewe

(4.79%). Although the share of respondents reporting English is above 47%, 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: it drops 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 3.2 reinforces this argument: 41.44% percent of the Akan group is found to be monolingual and 28.55% 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.

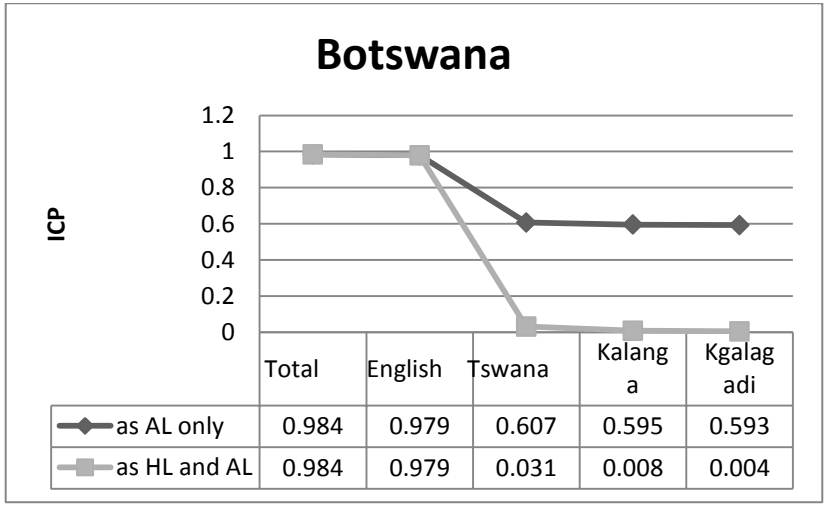
Figure 3.1 The drop in the communication potential by languages excluded as additional language (AL) only and both as home (HL) and additional language (AL)



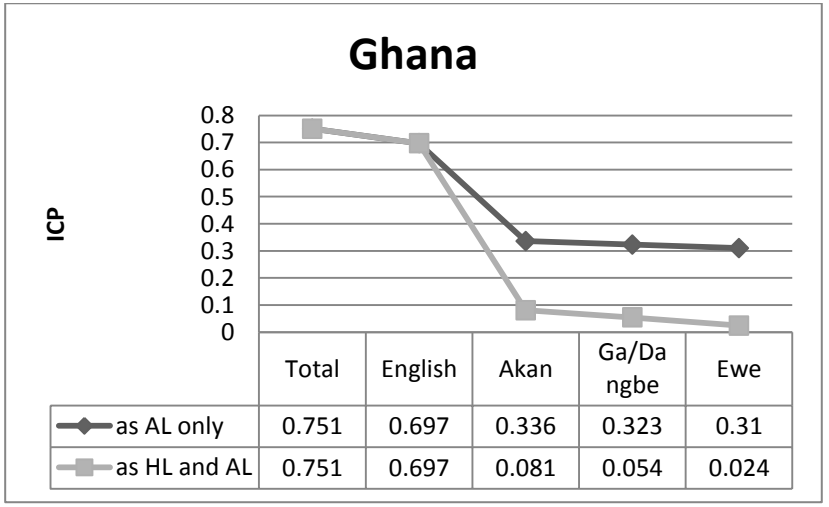
	Total	French	Fon	Yoruba	Dendi
as AL only	0.581	0.458	0.284	0.249	0.233
as HL and AL	0.581	0.458	0.184	0.132	0.114



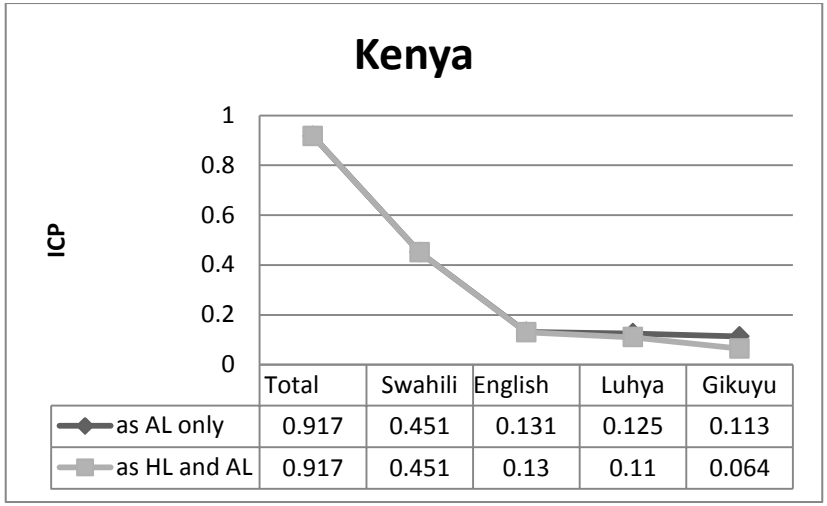
	Total	Dioula	French	Moore
as AL only	0.602	0.514	0.461	0.291
as HL and AL	0.602	0.514	0.459	0.026

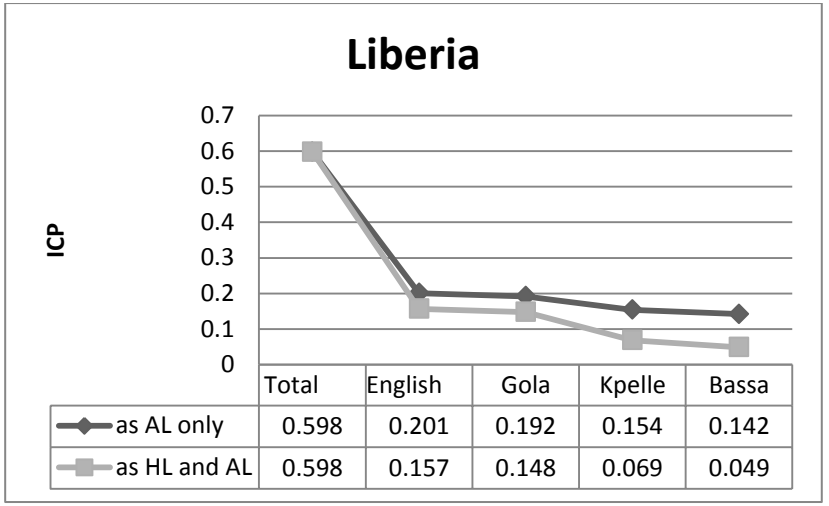


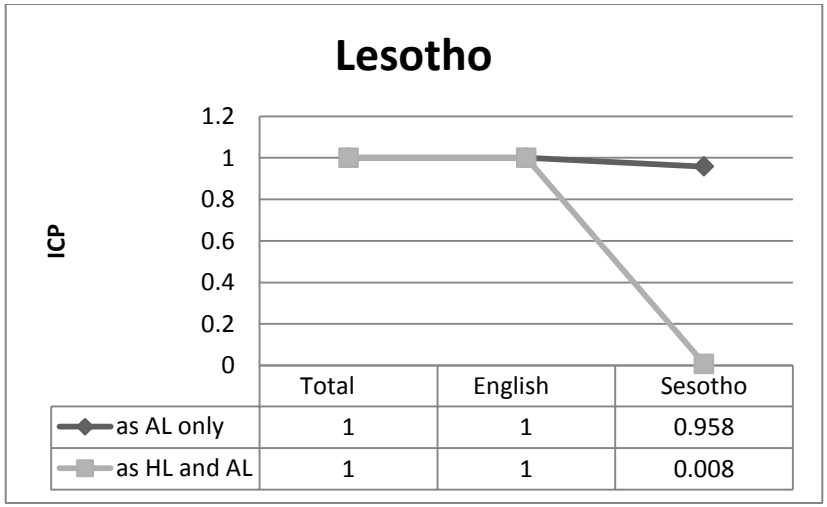
	Total	English	Tswana	Kalanga	Kgalagadi
as AL only	0.984	0.979	0.607	0.595	0.593
as HL and AL	0.984	0.979	0.031	0.008	0.004

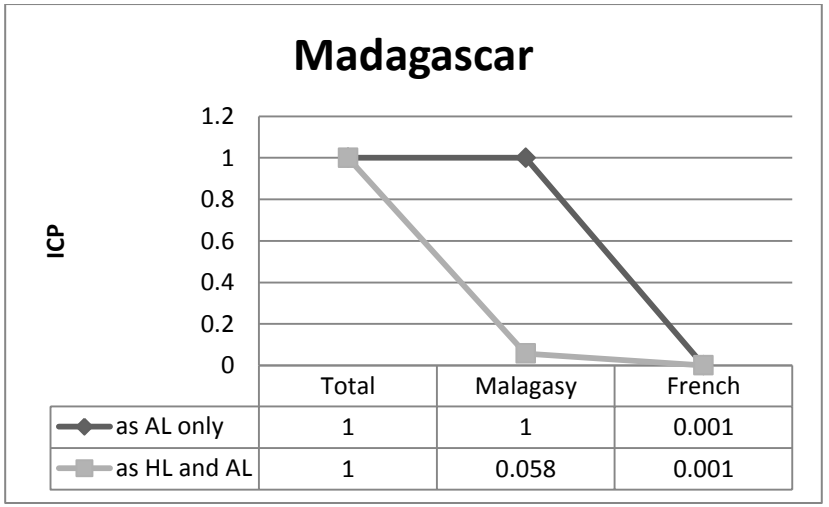


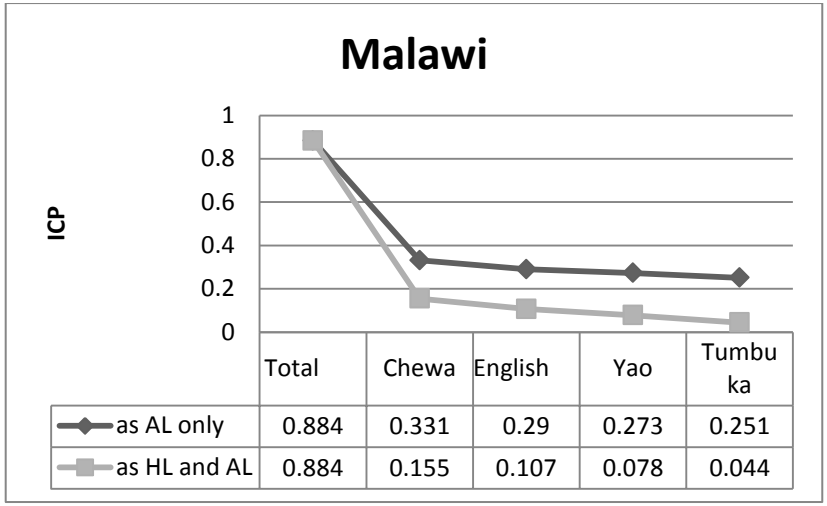
	Total	English	Akan	Ga/Dangbe	Ewe
as AL only	0.751	0.697	0.336	0.323	0.31
as HL and AL	0.751	0.697	0.081	0.054	0.024

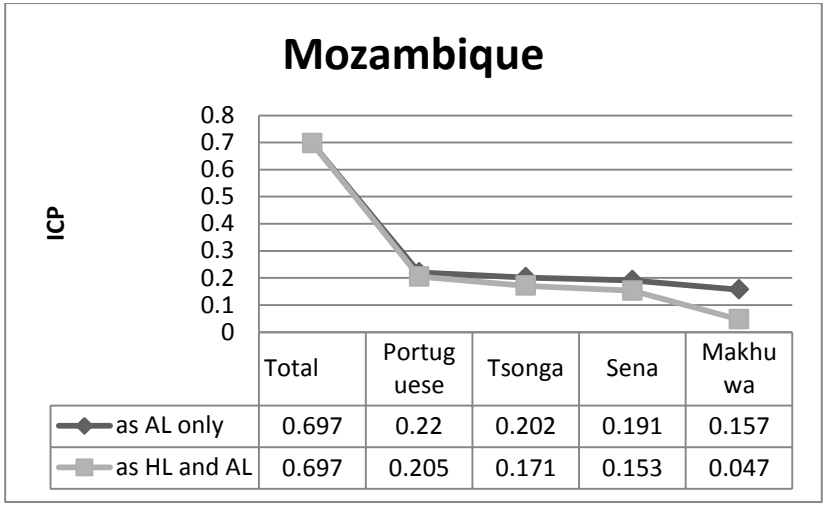

Kenya
ICP
1
0.8
0.6
0.4
0.2
0
Total Swahili English Luhya Gikuyu
as AL only 0.917 0.451 0.131 0.125 0.113
as HL and AL 0.917 0.451 0.13 0.11 0.064

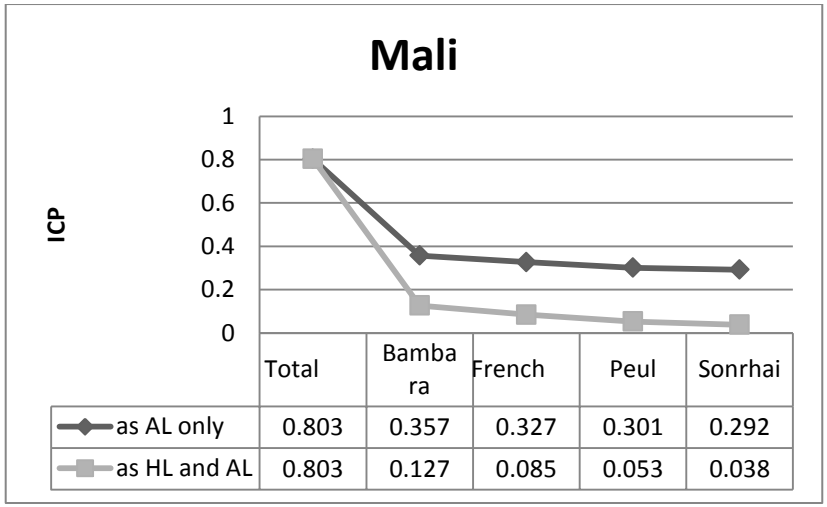

Liberia
ICP
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Total English Gola Kpelle Bassa
as AL only 0.598 0.201 0.192 0.154 0.142
as HL and AL 0.598 0.157 0.148 0.069 0.049

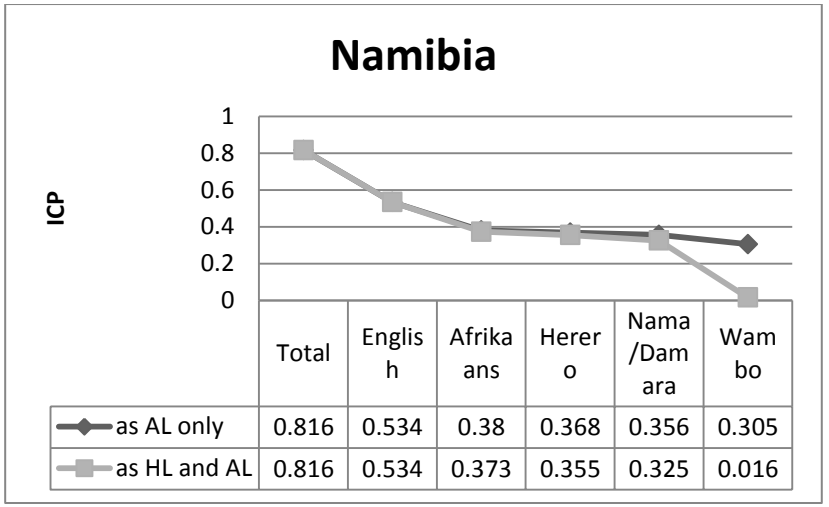

Lesotho
ICP
1.2
1
0.8
0.6
0.4
0.2
0
Total English Sesotho
as AL only 1 1 0.958
as HL and AL 1 1 0.008

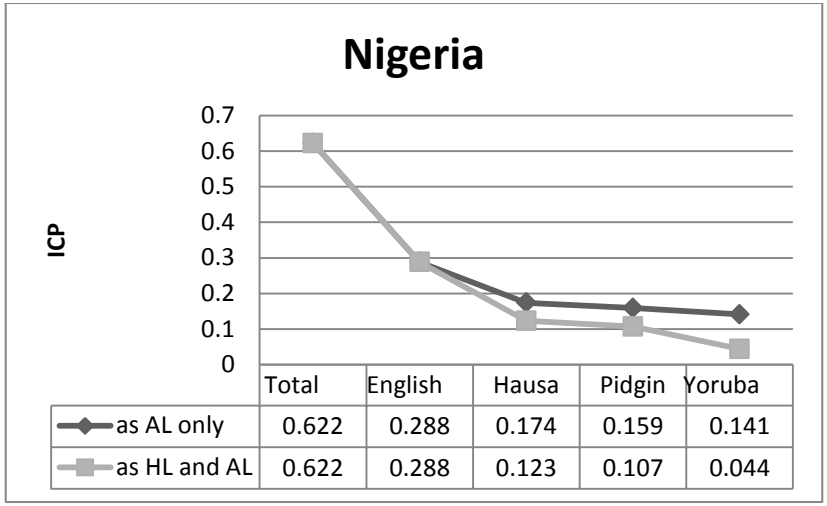

Madagascar
ICP
1.2
1
0.8
0.6
0.4
0.2
0
Total Malagasy French
as AL only 1 1 0.001
as HL and AL 1 0.058 0.001


Malawi
ICP
1
0.8
0.6
0.4
0.2
0
Total
Chewa
English
Yao
Tumbuka
as AL only
0.884
0.331
0.29
0.273
0.251
as HL and AL
0.884
0.155
0.107
0.078
0.044

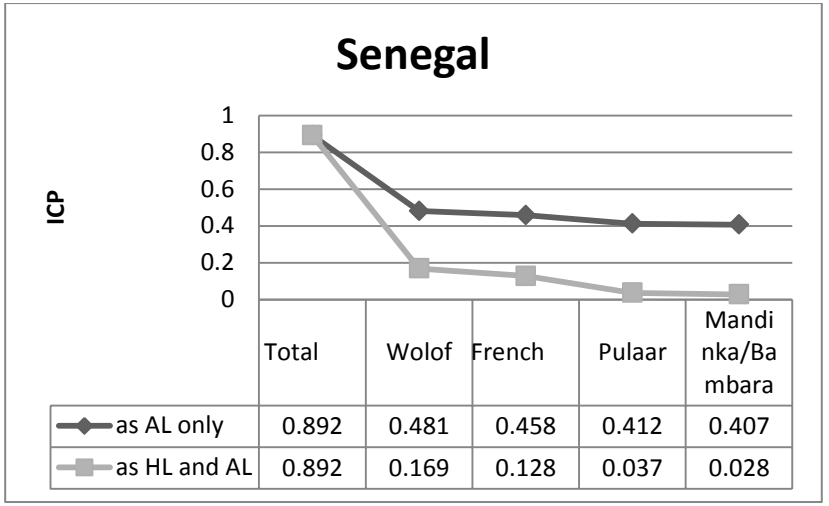

Mozambique
ICP
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Total
Portuguese
Tsonga
Sena
Makhuwa
as AL only
0.697
0.22
0.202
0.191
0.157
as HL and AL
0.697
0.205
0.171
0.153
0.047


Mali
ICP
1
0.8
0.6
0.4
0.2
0
Total
Bambara
French
Peul
Sonrhai
as AL only
0.803
0.357
0.327
0.301
0.292
as HL and AL
0.803
0.127
0.085
0.053
0.038

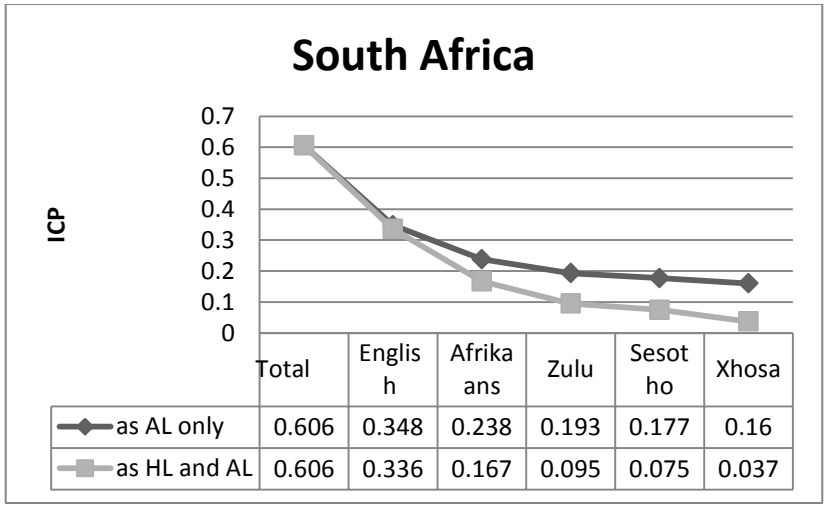

Namibia
ICP
1
0.8
0.6
0.4
0.2
0
Total
English
Afrikaans
Herero
Nama/Damara
Wambo
as AL only
0.816
0.534
0.38
0.368
0.356
0.305
as HL and AL
0.816
0.534
0.373
0.355
0.325
0.016



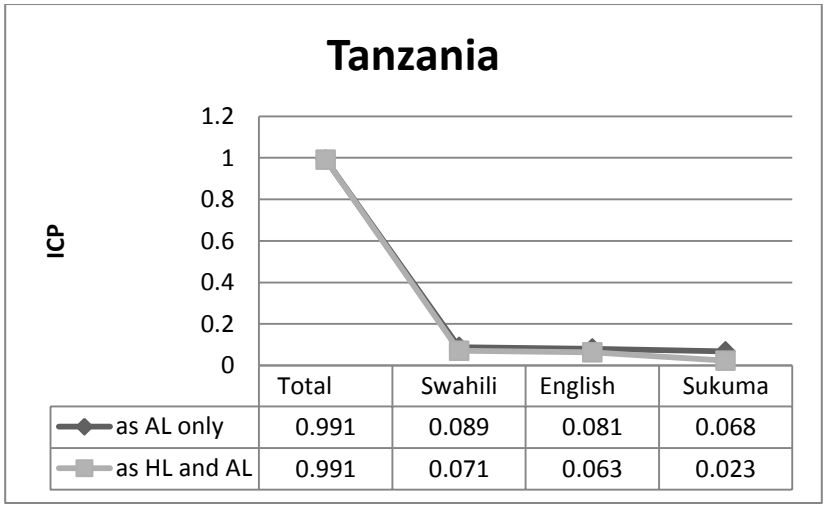
	Total	English	Hausa	Pidgin	Yoruba
as AL only	0.622	0.288	0.174	0.159	0.141
as HL and AL	0.622	0.288	0.123	0.107	0.044



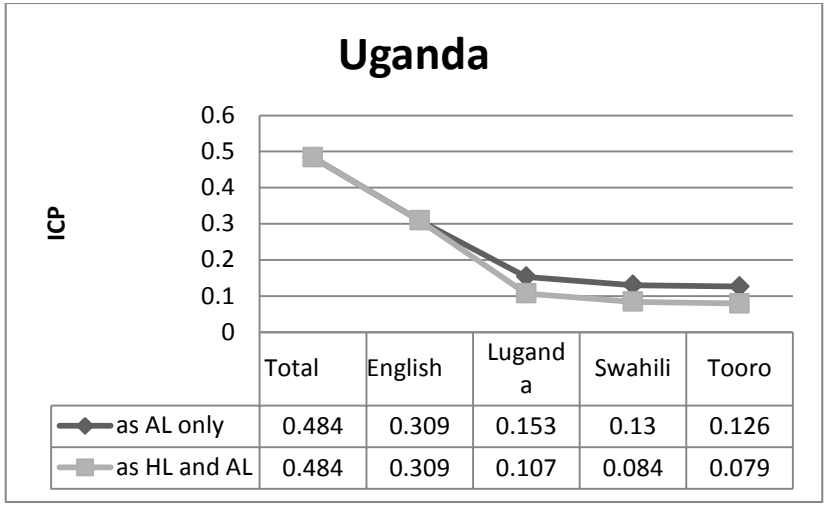
	Total	Wolof	French	Pulaar	Mandinka/Bambara
as AL only	0.892	0.481	0.458	0.412	0.407
as HL and AL	0.892	0.169	0.128	0.037	0.028

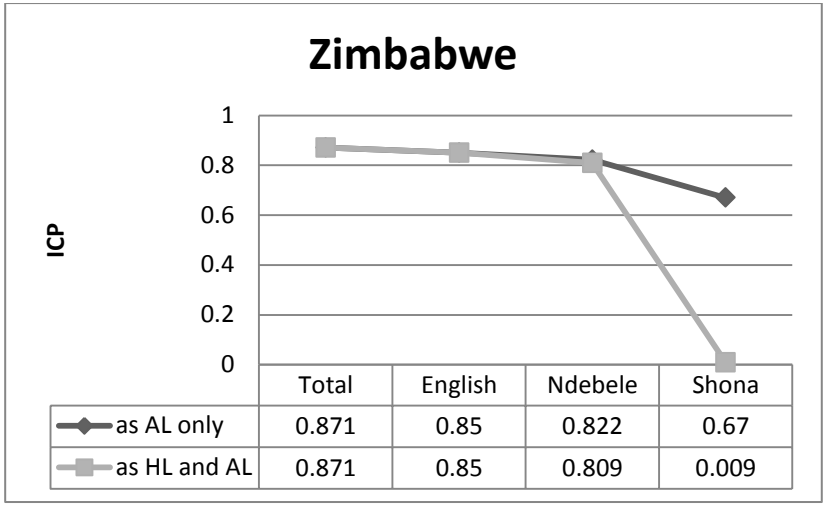


	Total	English	Afrikaans	Zulu	Sesotho	Xhosa
as AL only	0.606	0.348	0.238	0.193	0.177	0.16
as HL and AL	0.606	0.336	0.167	0.095	0.075	0.037



	Total	Swahili	English	Sukuma
as AL only	0.991	0.089	0.081	0.068
as HL and AL	0.991	0.071	0.063	0.023


Uganda
ICP
0.6
0.5
0.4
0.3
0.2
0.1
0
Total
English
Lugand a
Swahili
Tooro
as AL only
0.484
0.309
0.153
0.13
0.126
as HL and AL
0.484
0.309
0.107
0.084
0.079


Zimbabwe
ICP
1
0.8
0.6
0.4
0.2
0
Total
English
Ndebele
Shona
as AL only
0.871
0.85
0.822
0.67
as HL and AL
0.871
0.85
0.809
0.009

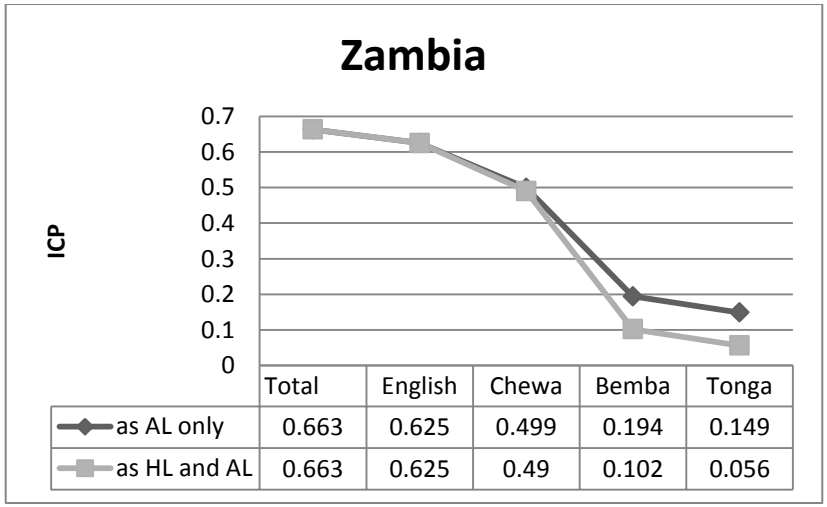

Zambia
ICP
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Total
English
Chewa
Bemba
Tonga
as AL only
0.663
0.625
0.499
0.194
0.149
as HL and AL
0.663
0.625
0.49
0.102
0.056

Table 3.2 The linguistic repertoire of people speaking Akan at home

home	additional	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 speaks Nzema, 3.1% speaks Hausa, 3.1% speaks Sehwi, 2.77 speaks French.

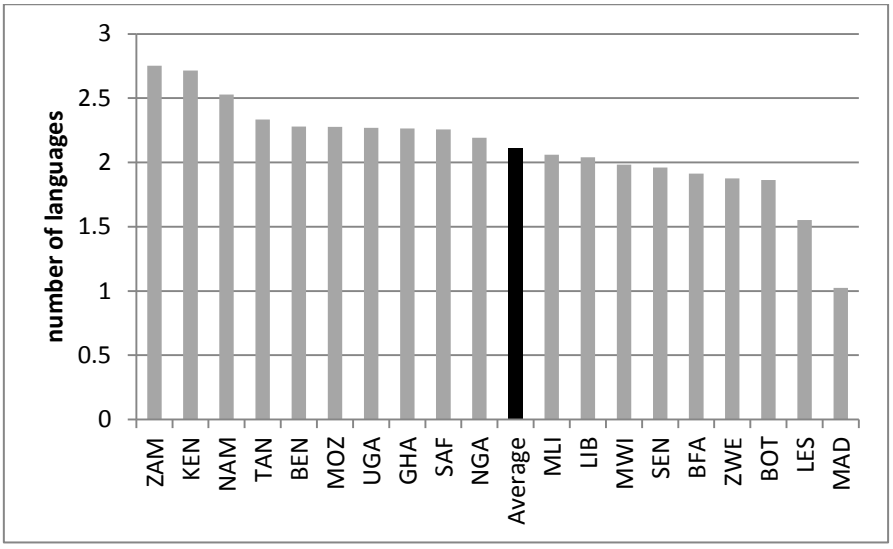
At this point, we remark that if languages are ordered according to a different aspect such as their size as home language or their total number of speakers, figures highlight other dimensions of the linguistic situation. If Akan was the first language to be omitted and English only afterwards, it would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

In order to make it more straightforward how to interpret the graphic representation of the ICP, we discuss the case of Liberia where the language situation is significantly different from that in Ghana (see Figure 3.1 and the supplementary material). The colonizer's language is selected as home language by 23.43% of the sample. None of the local languages are spoken by more than 10% as an additional language. The order of languages according to their reported frequency in Q88E is English (48.96%), Gola (9.61%), Kpelle (8.52%), and Bassa (5.48%). 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 grey lines suggests that a significant share of people speaking English the most often at home are 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 3.1 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 grey 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 3.2 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 2. Since, more than the half of the population is proficient in English along the home language and Swahili, Kenya is ranked second in Figure 3.2. Observing Figure 3.1 and 3.2, we can arrive at the following rule of thumb: the number of languages in the typical repertoire is above the average in countries where the light grey line does not drop much below 0.1 after the third language is excluded.

Figure 3.2 The number of languages in the typical repertoire per country



Note: Figure 2 presents weighted average. Sample weights are obtained from Afrobarometer.

The major benefit of the graphic representation is that it is applicable to analyze the relation between indigenous and European languages. Although in 12 out of the 19 countries the former colonizer'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 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 grey 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% as home language in the latter country.

The supplementary material and the graphic representation allow us to classify countries by their language use patterns. According to the distribution of the major indigenous languages, the sample countries can be organized into the following 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, we find Botswana, Ghana, Malawi, Mali, Senegal, and Zimbabwe, where the largest indigenous language is spoken by between 80 and 100 percent of the population but by only about 50 to 80 percent 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 50-60 percent, there are not any serious indigenous competitors. In the fourth group of countries, 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 characterized by a few relatively large regionally dominant languages.

Adding the languages of the former colonizers to the patterns we have just explained, the picture on the linguistic situation becomes even more sophisticated. As Figure 3.3 and 3.4 present, the distribution of European languages is dependent on the distribution of vernaculars. The share of the population proficient in the former colonizer'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 3.4). The latter two groups of countries in the above introduced classification scheme are scattered roughly in the upper left part of Figure 3 and 4. These countries are also the ones where the exclusion of the former colonizer's language results in a considerable drop in the average communication potential in Figure 3.1. Thus, the language of the former colonizer is the most widely spoken and the most important in terms of the communication potential in countries where there is not any indigenous language that could serve as a national lingua franca. However, without undermining the validity of this general pattern, there might be differences between countries formerly colonized 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 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 3.3, were included in the Afrobarometer, we had 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 3.3. Moreover, our 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.

Figure 3.3 The relationship between the size of the largest home language and the share of the population speaking the former colonizer's language

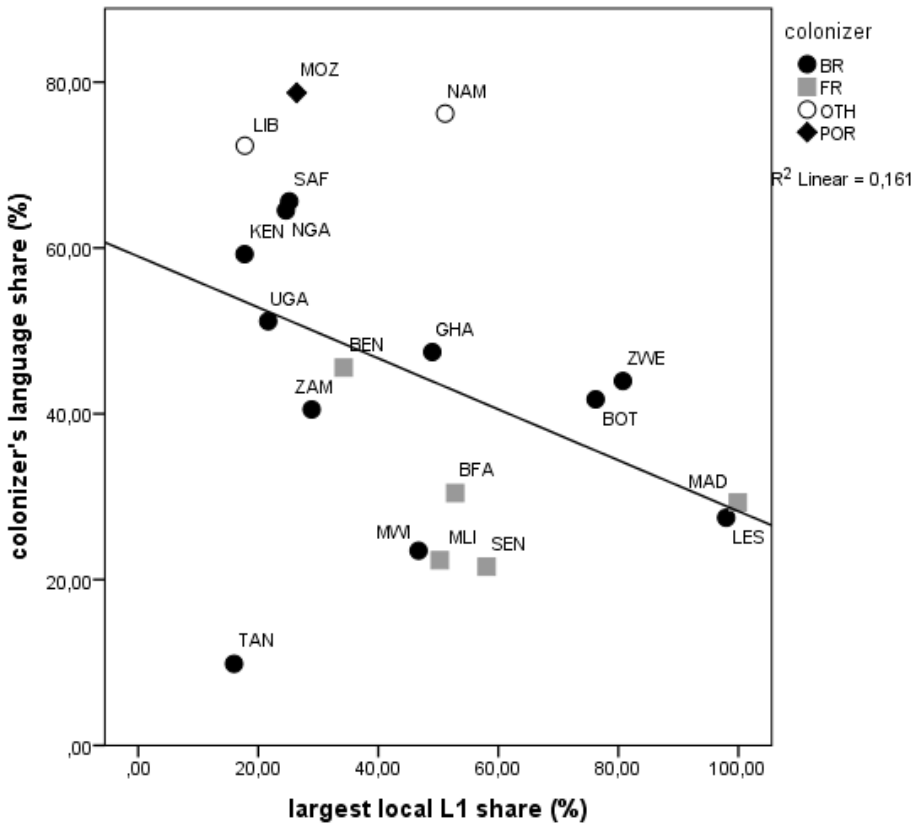
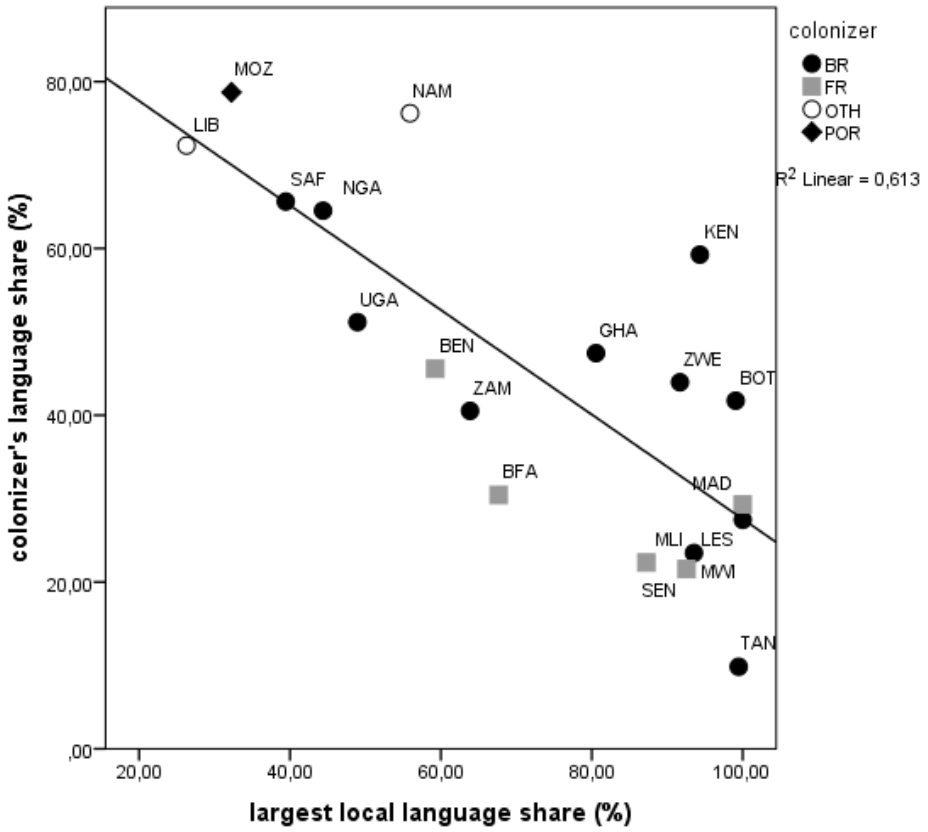


Figure 3.4 The relationship between the size of the most widely spoken and the former colonizer's language



3.6 Conclusion

Utilizing the 4th round of the Afrobarometer Survey, this study presents the most important dimensions of the linguistic situation in twenty Sub-Saharan African countries. Without repeating what has already been discussed in the previous sections, the conclusion is devoted to illustrate the relevance of our 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 Section 3.4, we argue 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, we identify the following five survey design-related factors that influence the observed linguistic situation: the detailedness of the classification scheme in the questionnaire, the data collection method, the properties of the investigated population, the purpose and the conceptual framework of the survey, and the respondents' behavior. We suggest that the above listed 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 chapter also indicates that taking other than first and home languages into account makes it possible to analyze some aspects of the linguistic situation that have gained only marginal attention so far. Since, as Table 3.1 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, we suggest 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 in the main text we refrained from 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 programs and to monitor language dynamics.

Appendix 3A

The construction of the Index of Communication Potential (ICP)

The basis of the individual and country level ICPs is a $n \times n$ symmetric matrix M_k (Eq. 3A.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 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.

$$\mathbf{M_k} = \begin{bmatrix} & 1 & 2 & \cdots & j & \cdots & n_k \\ 1 & m_{11k} & m_{12k} & \cdots & m_{1jk} & \cdots & m_{1n_k k} \\ 2 & m_{21k} & m_{22k} & \cdots & m_{2jk} & \cdots & m_{2n_k k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots \\ i & m_{i1k} & m_{i2k} & \cdots & m_{ijk} & \cdots & m_{in_k k} \\ \vdots & \vdots & \vdots & & \vdots & \ddots & \vdots \\ n_k & m_{n_k 1k} & m_{n_k 2k} & \cdots & m_{n_k jk} & \cdots & m_{n_k n_k k} \end{bmatrix} \quad \text{(Eq. 3A.1)}$$

The communication potential of individual *i* in country *k* is computed as shown in Eq. 3A.2.

$$\mathrm{icp_{ik}} = \sum_{\mathrm{j=1,j^1i}}^{\mathrm{n_k}} \mathrm{w_{jk} m_{ijk}} / (\sum_{\mathrm{j=1}}^{\mathrm{n_k}} \mathrm{w_{jk}} - \mathrm{w_{ik}}) = \sum_{\mathrm{j=1,j^1i}}^{\mathrm{n_k}} \mathrm{w_{jk} m_{ijk}} / (\mathrm{n_k} - \mathrm{w_{ik}}) \qquad \text{(Eq. 3A.2)}$$

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 to not take one's communication potential with oneself into account. The $\mathrm{icp_{ik}}$ 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 (Eq. 3A.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.

$$\mathrm{ICP_k} = (\sum_{\mathrm{i=1}}^{\mathrm{n_k}} \mathrm{w_{ik} icp_{ik}}) / \mathrm{n_k} \qquad \text{(Eq. 3A.3)}$$

Appendix 3B

Ethnic and linguistic diversity

Ethnic and linguistic diversity D_k in country k is computed using Eq. 3B.1.

$$D_k = 1 - \sum_{g=1}^{G_k} s_{gk}^2 \,, \qquad \text{(Eq. 3B.1)}$$

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. Eq. 3B.1 is also known as 1 minus the Herfindahl-index of concentration (Herfindahl 1950).

Supplementary material

Introduction

The aim of this supplementary material is to present available data on indigenous and European languages in the twenty sample countries and to compare them with the Afrobarometer (hereafter AB).

The linguistic situation of each country is described in a table. Reported values are understood as percentages. ***Column 1*** lists the name of groups as used and spelled in the AB codebook. The classification and the names of ethnic and linguistic groups in Q79 and Q3 are usually identical. Exceptions, if exist, are explained in the **General notes** under the tables. ***Column 2 and 3*** show the share of the listed ethnic and linguistic groups based on Q79 (on ethnicity) and Q3 (on home language), respectively. ***Column 4*** presents the share of the population speaking the listed languages as additional languages from Q88E. The total share of speakers is computed in ***Column 5***. In order to cross-check the Afrobarometer data, the remaining columns (***Column 6 to 11***) present linguistic information from Ethnologue (Lewis et al. 2014), the latest available national censuses, and other sources. However, census questionnaires are not standardised across countries: certain countries collect information on ethnicity, while others on home languages or both. The column headings make it clear which one of the two is reported.

The ***General notes*** have two additional aims. First, they list some references which contain information on the historical origins of the language situation, the spread of languages and their use in various domains (education, media etc.) and the design of language policies. The majority of references are published in two sources. Simpson (2008) contains comprehensive description of the linguistic and language policy situation in Senegal, Mali, Ghana, Nigeria, Kenya, Tanzania, Zambia and South Africa (and some other countries that are not included in the AB). The Current Issues in Language Planning and the Journal of Multilingual and Multicultural Development (both are published by Taylor and Francis) have published works on Botswana, Malawi, Nigeria and South Africa which were later republished in Baldauf and Kaplan (2004) and Kaplan and Baldauf (2007). Although there are a huge number of sources that focus on certain languages or regions within each country, we do not list them as general reference. Second, the General notes are devoted to draw attention to the special properties of a country's linguistic situation which eventually support the interpretation of the numbers displayed in the tables.

If additional information is available on individual languages, it is presented in ***footnotes*** under the general notes. Most data in the footnotes are based on literacy surveys and the following sources collected by Albaugh (2014 and 2012): Adegbija 2007 and 1994, Anyidoho and Dakubu 2008, Baker and Jones 1998, Benson 2010,

Bunyi 2007, Canvin 2007, Crystal 2003, Graddol 1997, Heugh 2007, Lewis 2009, Matiki 2006, McLaughlin 2008, Nkosana 2008, Nyika 2008, Leclerc, OIF 2007, Pawlikova-Vilhanova 1996, Skattum 2008.

Moreover, we report the official language(s), the ethnic and linguistic diversity indicators, the average Index of Communication Potential based on the Afrobarometer, and the number of living languages in the 17th edition of Ethnologue (Lewis et al. 2014) for each country.

Benin

Official language: French
Ethnic diversity: 0.825
Linguistic diversity: 0.816
Index of Communication Potential: 0.581
The number of languages in Ethnologue: 54

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue) [1]	(7) share as sociolinguistic group (census 2002)[2]
Fon[3]	31.75	34.24	25.02	59.26	27.53 (Fon: 17.02 Goun: 3.89)	39.2 (Fon: 17.6, Goun: 6.3)
Adja	16.88	15.43	9.35	24.78	12.89 (Adja: 4.38)	15.2 (Adja: 8.7)
Yoruba	13.78	13.33	11.4	24.73	13.32 (Yoruba: 5.65)	12.3 (Yoruba: 1.8)
Bariba	11.87	11.29	6.89	18.18	6.81 (Bariba: 5.59)	9.2
Goun	7.03	6.93	-	6.93	3.89	6.3
Otamari	6.49	6.65	1.28	7.93	5.3 (Otamari: 1.46)	6.1
Dendi	3.31	4.66	11.39	16.05	0.36	2.5
Yoa	3.20	2.89	1.1	3.99	5.72 (Yoa: 0.6)	4.5
French[4]	-	0.11	45.46	45.57	0.2	-

[1]In Column 6, we follow the concept of the 2002 census and present the share of the broader sociolinguistic group defined in the census and in Footnote 2. Thus, the numbers are more comparable across sources. The largest individual languages within the sociolinguistic groups are presented in parentheses.

[2]Source: INSAE (2003). The census reports the share of the sociolinguistic groups. The Fon group includes Fon, Goun, Aizo, Mahi, Oueme, Torri, Kotafon, Tofin, and Seto. The Adja group includes Adja, Sahoue, Xwla, Mina, Houedah, Ouatchi, and Defi. The Yoruba group in includes Nagot, Yoruba, Idaasha, Holli-Dje, Ife, Mokole, and Chabe. The Bariba group includes Bariba, Boo, and Boko. The Otamari group includes Berba, Ditamari, Waama, Natimba, Otamari, Gourmantche, Yende, Betyobe, and Gagamba. The Yoa group includes Yoa, Lokpa, Anii, Koto-Koli, Windji-Windji, Kabye, Soruba Biyobe, and Taneka. The Dendi group includes Dendi and Djerma.

[3]The share of the population speaking Fon-Ewe languages is about 60% in Adegbija (1994, p. 8).

[4]The share of the population speaking French in OIF (2007) is 8.8%, and according to the Organisation Internationale de la Francophonie (2010) is 32.39%. Based on the 2nd and 3rd census, the share of the population reporting to read and write in French and understanding it was 22.9% (as a share of the population aged 3 or above) in 1992 and

34.1% (as a share of the population aged 6 or above) in 2002 (Amadou Sanni and Atodjinou 2012, p. 39).

Botswana

Official languages: English, Tswana
Ethnic diversity: 0.923 (groups within the Tswana are distinguished)
Linguistic diversity: 0.407 (Tswana is treated as a single language, dialects are not distinguished)
Index of Communication Potential: 0.984
The number of languages in Ethnologue: 29

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)[4]	(7) share as L1 (demographic survey 2006)[1]
Setswana[2]	-	76.21	22.86	99.07	52.84 (7.41 as L2)	72.6
Kgatla	9.12	-	-	-	-	-
Kwena	9.16	-	-	-	-	-
Gwato	12.73	-	-	-	-	-
Ngwaketse	8.8	-	-	-	-	-
Rolong	3.14	-	-	-	-	-
Hurutse	3.68	-	-	-	-	-
Lete	4.1	-	-	-	-	-
Tswapong	5.14	1.5	0.98	2.48	0	2.2
Sarwa	3.1	1.67	1.45	3.12	0	2.2
Kalanga	14.45	8.22	7.07	15.29	7.41	8.4
Kgalagadi	7.04	4.07	3.11	7.18	1.98	2.9
Yeyi	2.38	0.8	0.4	1.2	0.99	0.5
Birwa	3.59	2.01	0.44	2.45	0.74	2.6
Mbukushu	2.43	2.04	0.85	2.89	0.99	2
English[3]	-	-	41.75	41.75	0.002 (31.11 as L2)	2.1

General notes: Information on the dimensions of the language situation including the use of language in public affairs, education and media is provided in Nyati-Ramahobo (2000) and Andersson and Janson (1997). The dialects of Tswana (Kgatla, Kwena, Lete, Ngwaketse, Ngwato, Rolong, Hurutse, Tawana, Thlaping, Tlokwa) are reported as separate ethnic groups in Q79 in the AB.

[1]Source: Central Statistics Office 2009. The reported percentages refer to the share of the population aged 2 or above that speak the listed languages as home language.

[2]The share of the population speaking Tswana is about 90% (80% as L1 and 10% as L2) in Obondo-Okoye and Sabone (1986), 93% in Baker and Jones (1998, p. 355) and 99% (90% as L1 and 9% as L2) in Adegbija (1994, p.11). In Leclerc (2009), 71.1% speaks Tswana as L1. According to the 2003 Literacy Survey (Central Statistics Office 2005, Table 38) surveying the population aged between 10 and 70 who never attended school or dropped out before Standard V, 3.5% has high and 82.9% has some writing competence, 17.4% has high and 54.6% has some reading competence, and 0.02% has high and 89.6% has some oral competence in Tswana.

[3] The share of the population speaking English is 35% in Nkosana (2008, p. 288), 38% in Graddol (1997, p. 11), and 40% in Crystal (2003, p. 62). According to the 2003

Literacy Survey (Central Statistics Office 2005, Table 38) surveying the population aged between 10 and 70 who never attended school or dropped out before Standard V, 38% has high and 41% has some writing competence, 34.2% has high and 36.1% has some reading competence, and 2.4% has high and 97.6% has some oral competence in English.

Burkina Faso

Official language: French
Ethnic diversity: 0.688
Linguistic diversity: 0.703
Index of Communication Potential: 0.602
The number of languages in Ethnologue: 69

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)[3]	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) as L1 (census 2006)[1]
Moore[2]	54.19	52.82	14.83	67.65	30.38	50.5
Fulfulde	6.47	6.63	2.49	9.12	4.56	9.3
Bissa	5.97	6.12	0.85	6.97	2.13	3.2
Bobo	6.02	5.37	1.36	6.74	1.83	Bobo: 1.4 Bwamu: 2.1
Gourounsi (Grusi)	5.04	4.39	1.89	6.28	3.19	at least 4
Gourmanchema	3.81	3.8	1.34	5.14	3.65	6.1
Dioula	0.74	3.74	32.18	35.92	6.08 (18.23 as L2 speaker)	4.9
Samo	3.28	2.82	0.78	3.60	1.39	1.9
Dagari	2.48	2.10	0.74	2.84	2.36	2.0
Marka	2.01	1.89	0.33	2.21	1.22	1.1
French[3]	-	1.42	29.02	30.44	-	1.3

General notes: The Bobo group in Column 6 includes Konabere, Bomu and Buamu. The Samo group in Column 6 includes the Samo-Matya, Samo-Maya, and Samo-Southern. According to Ethnologue (Lewis et al. 2014), Grusi (Gourounsi, Gurunsi), when used for languages, refers to Lyele, but mostly applies for the wider ethno-cultural group. The share of the Gourounsi group in Column 6 is based on the size of the associated groups, namely Kalamsé, Kasem, Lyélé, Northern Nuni, Southern Nuni, Pana, Phuie, Sissala and Winyé. The Gourounsi in Column 7 includes Lyele, Nuni, Kassena, Gourounsi, and Ko (Winyé).

[1]Source: Institut National de la Statistique et de la Demographie (2009a). The shares are based on the population aged 3 or above.

[2]The share of people speaking Moore is above 50% in Baker and Jones (1998, p. 356) and 50% in Adegbija (1994, p. 6).

[3]The share of population speaking French is 10 % in Baker and Jones (1998, p. 356), 5% in OIF (2007, p. 17), 20% in Leclerc (2009), and 19.62% according to OIF (2010). According to the 2006 Population Census (Institut National de la Statistique et de la Demographie 2009b), 3.06% of the population aged 10 or above is literate in French.

Cape Verde

Cape Verde is a unique country in Africa from several aspects. Before the arrival of Portuguese in the 15th century, the islands were uninhabited. Then, African slaves were brought here to work on Portuguese plantations. Today the islands are inhabited predominantly by mulattos. Ethnic groups as understood in other African countries are not known. Thus, Q79 in the Afrobarometer that is generally refer to the respondent's tribe or ethnic group is phrased differently in the Cape Verdean questionnaire: '*We have spoken to many [people in this country] and they have all described themselves in different ways. Some people describe themselves in terms of their language, religion, race, and others describe themselves in economic terms, such as working class, middle class, or a farmer. Besides being a [citizen of this country], which specific group do you feel you belong the first and foremost?*' Possible answers*: (0) Can't explain, (1) Language/tribe/ethnic group, (2) Race, (3) Region, (4) Religion, (5) Occupation, (6) Class, (7) Gender, (8) Individual/personal, (10) Won't differentiate/national identity, (12) Traditional leader, (13) Political party identity, (14) Age-related, (15) African/West African/Pan African, (16) Island, (50) Portuguese, (51) American, (60) Family/relationship-based (e.g. wife, parent, widow, etc.) (61) Marginalized group (e.g. disables, etc.), (995) Other, (999) Don't know*.

According to Q3 of the Afrobarometer, 0.32% speaks Portuguese and 99.68% speaks Cape Verdean Creole (Kabuverdianu or Krioulo) as home language. Unfortunately, second languages are not surveyed.

The ICP based on home languages only is 0.995. According to Leclerc (2011) 98% of the population speaks Capeverdian.

Ghana

Official languages: English
Ethnic diversity: 0.755
Linguistic diversity: 0.718
Index of Communication Potential: 0.751
The number of languages in Ethnologue: 81

(1) language/ ethnicity	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as ethnicity (census 2010)[1]
Akan[6]	45.36	48.99	31.57	80.56	33.66 (4.06 as L2)	47.5
Ewe	15.08	14.69	4.79	19.48	9.12 (2.03 as L2)	13.9
Ga/Dangbe	9.72	9.67	11.04	20.71	5.68	7.4
Dagbani	6.98	8.3	2.19	10.49	3.24	16.6
Dagari	2.26	3.28	0.45	3.73	2.84	-
English[7]	-	0.06	47.39	47.45	no data on L1 speakers (4.06 as L2)	-

General notes: General information on the dimensions of the language situation including the use of languages in public affairs, education and media is found in Anyidoho and Kropp Nakubu (2008). The Ga-Dangme people are considered as a

single ethnic group that speak different languages in Ethnologue (Lewis et al. 2014). However, Q3 on home languages in the Afrobarometer applies the same classification system as Q79 on ethnicity. Thus, unfortunately, we cannot distinguish between Ga and Dangbe as home languages. The linguistic diversity index is computed based on the classification scheme provided in Q3 (Ga and Dangbe are not distinguished). Dangbe is spelled 'Dangme' in Ethnologue (Lewis et al. 2014). The Akan people and language includes Twi, Fanti, Asante, Akwapim, Akyem and other subgroups and dialects.

[1]Source: Ghana Statistical Service (2012)

[2]The share of people speaking Akan is more than 50% in Baker and Jones (1998, p. 360) and Anyidoho and Dakubu (2008, p. 152), 41% in Lewis (2009), and 40% in Adegbija (1994, p. 9).

[3]The share of people speaking English is 7% in Graddol (1997) and Crystal (2003, p. 62), 10% in Lewis (2009), 32% in the population census of 2000 and above 65% in the population census of 2010 (20.1% of the population 11 years and older is literate in English only, 7% in Ghanaian language only, 45.8% in English and Ghanaian language, 0.3% in French and English, 0.8% in English, French and Ghanaian language) in Ghana Statistical Service (2012). A person was considered literate if she/he could read and write a simple statement with understanding.

Kenya

Official languages: English, Swahili
Ethnic diversity: 0.890
Linguistic diversity: 0.892
Index of Communication Potential: 0.917
The number of languages in Ethnologue: 67

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as ethnicity (census 2009)[1]
Gikuyu	18.07	17.72	3.02	20.74	17.15	17.15
Luhya	13.64	13.55	3.19	16.74	13.52	13.83
Luo	13.33	13.42	1.82	15.25	10.47	10.48
Kalenjin	12.26	12.36	0.43	12.79	13.27	12.87
Kamba	8.89	9.11	1.33	10.44	10.08 (1.55 as L2)	10.08
Meru/ Embu	8.04 (0.55 Embu)	8.23	0.69	8.92	5.41	5.13
Kisii	7.11	7.18	0.46	7.64	5.71 (1.3 as L2)	5.71
Somali	4.34	4.24	0.49	4.73	6.18	6.18
Mijikenda	3.82	4.19	0.52	4.71	4.51	5.08
Masai/ Samburu	2.25	2.15	0.59	2.74	2.18	2.79
Taita (Dawida)	2.32	2.03	0.09	2.12	0.7	0.7
Swahili[2]	0.23	1.36	92.96	94.32	0.3	0.3
English[3]	-	0.31	58.95	59.26	0.1 (7 as L2)	-

General notes: General information on the language situation is provided in Githiora (2008). In Ethnologue (Lewis et al. 2014), Kalenjin is a macrolanguage including Keiyo, Kipsigis, Markweeta, Nandi, Okiek, Pokot, Sabaot, Terik and Tugen. Luhya (Oluluyia in Ethnologue) is also a macrolanguage including Lubukushu, Luidakho-Luisukha-Lutirichi, Lukabaras, Lulogooli, Lutachoni, Nyala, Olukhayo, Olumarachi, Olumarama, Olunyole, Olushisa, Olutsotso, Oluwanga, Saamia. Mijikenda is a large ethnic group including 9 groups (Kigyriama, Chichonyi-Chidzihana-Chikauma, Chidigo and Chiduruma). Thus, Mijikenda as L1 and L2 based on the Afrobarometer includes Gyriama and Digo. The share reported in Column 4 is based on the size of all the related Kalenjin, Luhya and Mijikenda groups. Ethnologue reports data from the Kenyan Population and Housing Census of 2009. The possible small differences between the shares reported in Column 6 and 7 are explained with the rounding applied in Ethnologue and the differences in how the Kalenjin, Luhya and Mijikenda are grouped in the Ethnologue and the census. Data on literacy in English and/or Swahili are not found in available censuses and the Demographic and Health Surveys. Other information on Kenyan languages is available in Whiteley (1974).

[1]Source: Kenya National Bureau of Statistics (2010).

[2]The share of the population speaking Swahili is nearly 70% according to Baker and Jones (1998, p. 361), between 60% and 70% according to Pawlikova-Vilhanova (1996,

p. 162), 65% (5% as L1 and 60% as L2) in Adegbija (1994, p. 8) and Githiora (2008, p. 245), and 75% in Bunyi (2007, p. 22). Leclerc (2010) reports only 39.4%.

[3]The share of the population speaking English is 15% in Bunyi (2007, p. 22), 'barely a quarter' in Nabea (2009, p. 122), 9% in Graddol (1997, p. 11), 16% in Baker and Jones (1998, p. 361) and 8.8% in Crystal (2003, p. 63).

Lesotho

Official languages: English, Southern Sotho (Sesotho)
Ethnic diversity: 0.888 (based on clan distinction within the Sotho group instead of distinction between ethnicities)
Linguistic diversity: 0.040
Index of Communication Potential: 1.000
The number of languages in Ethnologue: 5

(1) language/ ethnicity	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)
Southern Sotho[1]	-	97.96	2.04	100	93.45
Mofokeng	17.13	-	-	-	-
Motebele	16.78	-	-	-	-
Mokoena	16.45	-	-	-	-
Mohlakoana	9.84	-	-	-	-
Motaung	9.32	-	-	-	-
Mosiea	5.90	-	-	-	-
Motloung	4.71	-	-	-	-
Motlokoa	3.74	-	-	-	-
Motsoeneng	2.77	-	-	-	-
Lekholokoe	2.65	-	-	-	-
English[2]	-	1.3	26.17	27.47	no data on L1 speakers (26.4 as L2 speakers)

General notes: The national census in 2006 and the Demographic Survey in 2011 do not contain information on home language and/or clan membership. Although the Demographic and Health Surveys in 2004 and 2009 ask the respondents' home language, they are not presented in the reports that are available online. Literacy in English and Sesotho is surveyed and tested in the census of 2006, but unfortunately, literacy is not reported by language.

[1]The share of the population speaking Sesotho is 97% in Leclerc (2010) and 99% (95% as L1 and 4% as L2) in Adegbija (1994, p. 10).

[2]The share of people speaking English is 27% in Graddol (1997, p. 11) and 23% in Crystal (2003, p. 63).

Liberia

Official languages: English
Ethnic diversity: 0.888
Linguistic diversity: 0.885
Index of Communication Potential: 0.598
The number of languages in Ethnologue: 31

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as ethnicity (census 2008)[1]
Kpelle[2]	22.22	17.77	8.52	26.29	21.86	20.29
Grebo	12.24	8.15	3.33	11.48	7.62	10.03
Bassa	12.14	8.75	5.48	14.23	11.59	13.42
Mano	9.17	7.31	2.8	10.11	8.77	7.87
Gio (Dan)	7.55	6.68	3.52	10.19	8.92	7.97
Lorma	7.24	4.47	2.59	7.06	5.61	5.13
Gola	7.00	5.09	9.61	14.7	2.86	4.4
Kru	5.85	4.31	2.37	6.68	6.13 (Klao)	6.04
Kissi	5.30	4.38	1.28	5.66	3.31	4.83
Krahn	3.28	2.86	0.5	3.36	3.34 (including Eastern and Western Krahn)	4.00
Vai	2.53	1.54	3.98	5.52	2.99	4.03
Manding	2.02	1.55	1.09	2.64	1.45	3.18
English[3]	-	23.43 (including simple Liberian English)	48.96 (including simple Liberian English)	72.34 (including simple Liberian English)	2.01 standard English (43.15 Liberian English as L2)	-

General notes: Although the population census surveys self-reported literacy, we do not have information on the respondents' reading and writing skills per language.

[1]Source: Liberia Institute of Statistics and Geo-Information Services (2009)

2The share of the population speaking Kpelle is 60% (20% as L1 and 40% as L2) in Adegbija (1994, p.10).

[3]The share of the population speaking English is 20% in Baker and Jones (1998, p. 362.), 91% in Graddol (1997, p. 11) and 90% (40% as L1 and 50% as L2) in Adegbija (1994, p. 10). I assume that these numbers include or predominantly refer to people speaking Liberian English which in is named Krio in Albaugh (2014) and Adegbija (1994). The Liberia Demographic and Health Survey 2013 () that aims to measure the most important characteristics of the population aged between 15 and 49 also surveys literacy. However, instead of relying on the self-reported skills, literacy among respondents with education below the secondary level is tested. People with higher than secondary schooling are assumed to be literate. Since most of the local languages have no accepted written script and are not taught in schools, and English is widely

spoken, the questionnaire is not translated into vernaculars. Thus, we assume that the literacy rate reported in the DHS reflects literacy in English. In the DHS, 55.11% of the surveyed population was literate. As a comparison, the 2008 population census (Liberia Institute of Statistics and Geo-Information Services 2009) finds that 55.86% of the population aged 10 or above is literate in any language.

Madagascar

Official languages: French, Malagasy (Plateau)
Ethnic diversity: 0.826 (groups within Malagasy as shown in the table are distinguished)
Linguistic diversity: 0.020
Index of Communication Potential: 1.000
The number of languages in Ethnologue: 18

(1) language/ ethnicity	(2) share as ethnicity (AB)[1]	(3) share as home language (AB)[2]	(4) share as additional language (AB)[3]	(5) total share (AB)	(6) share as L1 (Ethnologue)
Malagasy[1] (Merina, Plateau, national)	32.35	40.70	0.09	40.79	34.69
Malagasy (other dialect)	-	59.21	-	59.21	65.13
Betsileo	19.00	-	-	-	-
Betsimisaraka	15.05	-	-	-	13.05
Antemoro	5.10	-	-	-	-
Sakalava	4.94	-	-	-	1.61
Antesaka	3.35	-	-	-	5.21
Vezo	3.09	-	-	-	-
Antandroy	3.04	-	-	-	3.99
Antanosy	2.81	-	-	-	2.35
Sihanaka	2.31	-	-	-	-
Tsimihety	2.11	-	-	-	7.45
French[2]	-	0.09	29.21	29.3	0.001
English	-	-	6.2	6.2	-

General notes: The last column contains Northern and Southern Betsimisaraka. Betsileo and Sihanaka are Merina dialects. Vezo is a Sakalava dialect. Since they are considered mutually intelligible, the different versions of the Malagasy language that are listed as separate languages in Ethnologue (Lewis et al. 2014) are not distinguished when computing linguistic diversity.

[1]According to Baker and Jones (1998, p. 362) 98% of the population speaks Malagasy.

[2]The share of the population speaking French is 25% in Leclerc (2011), 5% in OIF (2007, p. 17), and 20% according to OIF (2010). Literacy in French is surveyed neither in the Demographic and Health Surveys nor in the population census of 1993.

Malawi

Official language: English
Ethnic diversity: 0.781
Linguistic diversity: 0.728
Index of Communication Potential: 0.884
The number of languages in Ethnologue: 16

(1) language/ ethnicity	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as tribe (census 2008)[1]	(8) share as L1 (Williams 1998)[2]	(9) share as L2 (Williams 1998)[2]
Chewa/Nyanja[3]	37.61	46.71	46.84	93.55	46.98	38.43	27	53
Lomwe	18.04	14.59	3.12	17.71	5.7	17.56	-	-
Tumbuka	11.34	11.64	6.33	17.97	14.76	8.84	11	4
Yao	11.46	9.93	7	16.93	20.91	13.51	19	1
Ngoni	11.53	7.77	0.78	8.55	-	11.46	-	-
Manganja	5.23	4.00	0.26	4.26	-	-	-	-
Sena	3.03	3.04	2.25	5.29	1.81	3.59	-	-
English[4]	-	0.09	23.4	23.49	0.001 (3.62 as L2)	-	-	-

General notes: Kayambazinthu (1998) and Lora-Kayambazinthu (2003) provide a comprehensive discussion on the dimensions of the Malawian language situation and the characteristics of the language policy. Although Ngoni and Manganja are listed as different ethnicities and languages in the AB questionnaire, according to Ethnologue Ngoni can be a name of a Chewa and also a Zulu dialect, Manganja is known as a dialect for Chewa and also Sena.

[1]Source: National Statistical Office (2008)

[2]Source: Williams (1998)

[3]The share of the population speaking Chewa is 76.6% (50.2% as L1) in Matiki (2006, p. 241) and Lora-Kayambazinthu (2003, p. 149) who use data from the 1966 population census, 75% in Baker and Jones (1998, p. 362), 60% (50% as L1 and 10% as L2) in Adegbija (1994, p. 7) and 50% in Leclerc (2011).

[4]The share of English speakers is 5% according to Graddol (1997, p. 11) and Crystal (2003, p. 63). According to the population census 2008 (National Statistical Office of Malawi), 24.07% of the population older than 5 years is literate in English. The corresponding proportion is 32% in the 2000 census (Lora-Kayambazinthu 2003, p. 151).

Mali

Official language: French
Ethnic diversity: 0.839
Linguistic diversity: 0.719
Index of Communication Potential: 0.803
The number of living languages in Ethnologue: 66

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as L1 (census 2009)[1]
Bambara[2]	31.81	50.25	36.98	87.23	27.54 (68.85 as L2)	46.5
Peul/ Fulfulde	15.94	7.91	12.58	20.49	8.09	9.39
Senoufo/ Mianka	10.57	7.63	2.72	10.35	8.68 (Mianka: 5.16)	4.29
Sonrhai	7.42	7.30	6.04	13.34	4.44	5.58
Dogon	7.28	6.90	0.67	7.57	3.4	7.12
Soninke/ Sarakolle	6.95	5.21	4.58	9.79	8.81	6.33
Malinke	7.07	4.52	2.77	7.29	13.47	5.6
French[3]	-	0.12	22.22	22.34	0.0006	0.0001

General notes: The language situation is discussed in details in Skattum (2008). Sarakolle is an alternate name for Soninke in Ethnologue (Lewis et al. 2014). Sonrhai includes to Songhay-Hamburi Senni, Songhay-Koya Chiini, and Songhay-Koyraboro Senni, which are listed as individual languages in Ethnologue (Lewis et al. 2014). The share of people speaking Sonrhai in Column 6 contains all the aforementioned three languages. Sonrhai as ethnic group is 5.85% of the population. Malinke might refer to Eastern and Western Maninkakan, Maninkakan-Kita, and Khassonke. Numbers in Column 6 contain all these languages. Peulh/Fulfulde includes Fulfulde, Maasina and Pulaar. In Ethnologue Minianka (Mamara) is an individual language among the Senoufo languages. Senoufo languages include Mamara, Shempire, Sicite, Supyire, and Syenara. Ethnologue lists 19 Dogon languages. Column 6 shows the total share of these groups within the country.

[1]Source: Institute National de la Statistique (2011). Since in the 2009 population census the Sonrhai and Djerma are considered as a single group, the share reported in Column 7 contains both of them. The Soninke is combined with the Maraka in the census and in Column 7. Since other Senoufo languages are not surveyed separately, the share of the Senoufo/Minianka group is completely based on the number of Minianka speakers in Column 7. Percentages are understood as the share of the population aged 6 or above. The census does not survey French as a home language separately, but provides the number of the population speaking a non-African (not Arabic) foreign language as mother tongue which is assumed to approximate the share of French as L1 in this study.

[2]The share of people speaking Bambara is 80% in Canvin (2007, p. 158) and Skattum (2008, p. 99), and 51% (31% as L1 and 20% as L2) in Adegbija (1994, p. 6).

[3]The share of the population speaking French is 5% in Baker and Johnson (1998, p. 363), 5-10% in Skattum (2008, p. 99), 8.2% according to the OIF (2007) and 18.23 in OIF 2(2010). Based on the population census (Institute National de la Statistique 2011), 22.56% of the population over 12 could read and write in French only and 2.32% in French and one of the national languages in 2009.

Mozambique

Official language: Portuguese
Ethnic diversity: 0.874 (Ethnic diversity is computed based on the sample that knows his ethnic group (1031 individuals instead of 1200. As the table shows the share of the sample that does not know his ethnic group is considerable (14.09%)).
Linguistic diversity: 0.872
Index of Communication Potential: 0.697
The number of languages in Ethnologue: 43

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as home language (census 2007)[1]
Makhuwa	25.07	26.40	5.85	32.25	23.71	25.34
Changana (Tsonga)	10.66	12.58	7.2	19.78	7.31	10.34
Lomwe	7.15	8.47	3.17	11.64	6.41	6.9
Sena	7.06	7.07	6.02	13.09	5.73	7.14
Nyungwe	3.96	5.16	1.95	7.11	1.88	2.75
Chuwabo	4.79	4.75	4.02	8.77	4.05	4.21
Ndau	4.13	4.24	4.93	9.17	6.75	4.14
Gitonga	2.70	2.84	1.75	4.59	1.6	1.13
Chewa/Nyanja	1.82	2.69	3.37	6.06	2.56	4.6
Portuguese[2]	-	14.44	64.29	78.72	5.73 as L1 (26.93 as L2)	12.76 (total speakers: 51.81)
Don't know	14.09	-	-	-	-	-

General notes: Lopes (1998) provides a complex overview on the language situation of Mozambique including the spread of Portuguese and indigenous languages, their use in education and media, and the features of the language policy. In Ethnologue (Lewis et al. 2014), Changana is a dialect of Tsonga and since the Afrobarometer questionnaire does not include Tsonga, we assume that the name of Changana is meant to refer to Tsonga. The share of Makhuwa in Column 6 is based on all Makhuwa varieties (Makhuwa, Makhuwa-Marrevone, Makhuwa-Meetto, Makhuwa-Moniga, Makhuwa-Saka, and Makhuwa-Shirima).

[1]Source: Instituto Nacional de Estatistica (2007). Percentages are understood as the share of the population aged 5 or above. The population census surveys mother tongue and the most frequently used language at home separately. Since there is only a slight difference between the two, we report the latter only.

[2]The share of the population speaking Portuguese is about 40% in Lopes (1998, p. 447), 33% (6% as L1+27% as L2) in Benson (2010, p. 238). Leclerc (2011) reports that 17% of the urban population speaks Portuguese, while the rural population is not likely to command it. Baker and Jones (1998, p. 364) argues that a quarter of Mozambicans are bilingual in Portuguese and an African language. According to the

Government Census of 2007 (Instituto Nacional de Estatistica 2007), the share of the population speaking Portuguese is 51.81%.

Namibia

Official language: English
Ethnic diversity: 0.705 (Nama and Damara are different groups)
Linguistic diversity: 0.701 (Nama and Damara are not separated)
Index of Communication Potential: 0.816
The number of languages in Ethnologue: 30

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as L1 (census 2011)[1]
Wambo[2]	51.04	51.15	4.75	55.9	51.43	48.9
Nama/Damara	14.15	13.39	5	18.39	9.47	11.3
Afrikaans	3.21	8.41	39.52	47.93	4.25	10.4
Herero	7.43	7.46	8.38	15.84	9.75	8.6
Kwangali	-	6.45	3.9	10.35	3.46	-
Lozi	1.42	3.46	1.53	4.99	1.35	-
Diriku	-	2.62	0.62	3.24	0.5	-
Kavango	10.39	-	-	-	-	8.5
English[3]	0.46	0.85	75.37	76.22	0.5 (14.2 as L2)	3.4

General notes: Pütz (1995) discusses various aspects of the Namibian language situation. Wambo includes Ndonga, Kwanyama and Kwambi according to Ethnologue (Lewis et al. 2014). In Q3 on home languages, Nama and Damara are not separated, while in Q79 on ethnicity they are listed separately.

[1]Source: Namibia Statistics Agency (2011). The percentages are understood as the share of households not individuals.

[2]According to Baker and Jones (1998, p. 364) the share of people speaking Ndonga (the biggest language within the Wambo group) is more than 50%.

[3]According to Graddol (1997, p. 11) 18% of the population speaks English. The corresponding share in Crystal (2003, p. 63) is 17%. According to the Population Census of 2001 (Namibia Statistics Agency 2003), 56.4% of the population above 15 is literate in English. However, reported ability was not tested.

Nigeria

Official language: English
Ethnic diversity: 0.856
Linguistic diversity: 0.876
Index of Communication Potential: 0.622
The number of languages in Ethnologue: 529

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as ethnicity (census 1963)[1]	(8) share as ethnicity (DHS 2008)[2]
Hausa[3]	24.43	24.63	19.74	44.37	12.32 (9.99 as L2)	20.9	22.3
Yoruba	22.09	22.14	4.28	26.42	12.58 (1.33 as L2)	20.3	17.7
Igbo	16.68	16.67	1.74	18.41	11.98	16.6	15.9
Ijaw	5.01	4.87	0.8	5.67	1.18	2	3.5
Fulani	4.03	4.10	2.33	6.43	5.07	8.6	6.1
Tiv	2.96	3.04	0.33	3.37	1.47	2.5	2.4
Ibibio	2.28	2.22	0.44	2.66	1 (3 as L2)	3.6	2.5
Edo	2.23	2.14	0.33	2.47	0.67	1.7	-
Kanuri	2.15	1.93	1.49	3.42	2.5	4.1	2
Pidgin English	-	0.03	13.82	13.85	19.97 including L1 and L2 speakers	-	-
English[4]	-	-	64.53	64.53	no data on L1 speakers (39.95 as L2)	-	-

General notes: The Nigerian language situation and the aspects of language policy are discussed in Adegbija (2007) and Simpson and Oyètádé (2008). The 1991 and 2006 population censuses do not survey ethnicity or home language, only nationality. Since the latest census that accounts for ethnicity is from 1963, we report the ethnic distribution from the Demographic and Health Survey 2008 surveying the population aged between 15 and 49. We took Izon from Ethnologue as Ijaw, and Adamawa Fulfulde as Fulani. Data on Kanuri from Ethnologue contain Central, Manga and Tumari groups.

[1]Source: Oshungade (1995)

[2]Source: National Population Commission (2009)

[3]According to Adegbija (1994, p. 11) 30% of the population speaks Hausa as first language and 20% as second language.

[4]The share of the population speaking English is between 20% and 30% according to Baker and Jones (1998, p. 365), less than 20% according to Adegbija (2007, p. 204), and 47% in Crystal (2003, p. 52 and p. 64) including Pidgin English. Simire (2004, p. 139) estimates that approximately 33 percent of the population can speak, read and write English. According to the National Literacy Survey 2010 (National Bureau of Statistics 2010), which is based on self-report, 57.9% of the adult population is literate in English. The literacy rate among young people (6-14 years) is 76.3%.

Senegal

Official language: French
Ethnic diversity: 0.701
Linguistic diversity: 0.605
Index of Communication Potential: 0.892
The number of languages in Ethnologue: 38

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 Ethnologue	(7) share as ethnicity (population census of 2002)[1]
Wolof[2]	45.37	58.06	34.5	92.56	39.36	42.01
Pulaar/ Toucouleur	25.72	21.01	11.07	32.08	27.44	26.55
Serer	14.14	9.20	5.28	14.48	11.32	14.82
Mandinka/ Bambara	6.62	5.64	6.54	12.18	6.7	3.71
Diola	4.91	4.02	3.26	7.28	3.4	4.3
French[3]	-	0.14	21.56	21.71	0.2	-

General notes: Further information on the Senegalese language situation in general and its relationship with national identity is found in McLaughlin (2008). The historical origins of multilingualism and the emergence of lingua francas are discussed in Mansour (1980).

[1]Source: Agence Nationale de la Statistique et de la Demographie (2006)

[2]The share of the population speaking Wolof is close to 90% according to McLaughlin (2008, p. 85) and Leclerc (2010) and 82% (42% as L1 and 40% as L2) in Adegbija (1994, p. 9).

[3]The share of the population speaking French is between 15% and 20% in Leclerc (2010), 10% according to OIF (2007), and 24.35% according to OIF (2010). According to the Population Census of 2002 (ANSD 2006), 37.8% of the population aged 6 or above is literate in French, 25.9% in Arabic, 1.5% in Wolof, and 1.2% in Pulaar.

South Africa

Official languages: Afrikaans, English, Ndebele, Northern Sotho (Pedi), Southern Sotho (Sesotho), Swati (Swazi), Tsonga (Shangaan), Tswana, Venda, Xhosa, Zulu
Ethnic diversity: 0.866
Linguistic diversity: 0.855
Index of Communication Potential: 0.606
The number of languages in Ethnologue: 31

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L2 (Ethnologue)	(7) share as L1 (Census 2011)[1]
Zulu	21.7	25.17	14.28	39.45	30.3	22.7
Xhosa	14.8	16.89	5.98	22.87	21.3	16.0
Afrikaans	8.4	15.34	16.37	31.71	19.9	13.5
Tswana	8.4	8.72	4.75	13.47	14.9	8.0
Pedi	6.6	8.07	3.49	11.56	17.6	9.1
Sotho	7.3	6.83	12.21	19.04	15.3	7.6
Tsonga (Shangaan)	3.8	3.83	4.75	8.58	6.6	4.5
Venda	1.8	2.72	1.43	4.15	3.3	2.4
Swati	1.9	2.31	2.34	4.65	4.6	2.6
Ndebele	1.2	1.15	1.47	2.62	2.7	2.1
English[2]	2.9	8.68	56.94	65.62	21.3	9.6
White	0.3	-	-	-	-	-
Couloured	4.5	-	-	-	-	-
Indian	2.2	-	-	-	-	-
South African only	13	-	-	-	-	-

General notes: Information on the South African language situation and the aspects of language policy are provided by UNESCO (2000), Kamwangamalu (2001) and Mesthrie (2008). Utilising population census data from 1980, 1991, and 2001, five linguistic atlases (Grabler et al. 1990, Krige 1994, Tait 1996, van der Merwe and van Niekerk 1994, van der Merwe and van der Merwe 2006) have been published. Since Ethnologue (Lewis et al. 2014) relies on the 2011 population census (Statistics South Africa 2012) data we do not report the share of L1 speakers from Ethnologue separately. The Ethnologue also reports the number of people that speak official languages as L2. Unfortunately, online census reports do not provide information on literacy per language.

[1]Source: Statistics South Africa (2012)

[2]According to Heugh (2007, p. 192), the share of the population speaking English as first language is up to 12%. The share of the population speaking English is 25% in Graddol (1997, p. 11) and 35.5% in Crystal (2003, p. 64).

Tanzania

Official language: Swahili
Ethnic diversity: 0.954
Linguistic diversity: 0.950
Index of Communication Potential: 0.991
The number of languages in Ethnologue: 129

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as L1 (Language Atlas, 2009)[1]	(8) share as L2+L3+… (Language Atlas, 2009)[1]	(9) share as L1 (census 1957 and Molnos 1967)[2]
Sukuma	15.59	15.96	4.25	20.21	12.09	12.98	2.1	8.88
Swahili[3]	-	8.91	90.55	99.46	33.37 (80% of rural population as L2)	2.04	4.87	0.1
Ha	4.66	4.66	3.74	8.4	2.2	2.74	0.8	2.35
Nyamwezi	4.16	4.43	0.73	5.16	2.18	2.1	0.7	2.95
Gogo	4.09	4.09	0.52	4.61	3.21	2.44	0.5	2.43
Iraqw	4.00	3.9	0.13	4.03	1.03	1.49	0.3	1.1
Jita	2.76	2.64	0.25	2.89	0.46	0.5	0.6	0.7
Makonde	3.14	2.59	1.33	3.92	2.18	1.86	0.5	2.71
Pare	3.03	2.57	0.77	3.34	1.11	1.18	0.4	1.02
Chaga	3.10	2.54	0.68	3.22	1.33	1.99	0.9	2.58
Haya	2.49	2.24	1.73	3.97	2.89	2.12	0.3	2.64
Nyaturu	2.04	2.16	0.52	2.68	1.32	1.42	0.2	1.59
Bena	2.02	2.02	0.29	2.31	1.49	1.11	0.6	1.59
English[4]	-	-	9.85	9.85	no data on L1 speakers (8.9 as L2)	-	-	-

General notes: Additional information on the linguistic situation and the aspects of language policy in Tanzania are found in Topan (2008), Rubagumya (1990), Polome and Hill (1980), and Molnos (1969).

[1]Source: Chuo Kikuu cha Dar es Salaam (2009). Values are understood as the share of population in 2002 (34443603).

[2]Since the 1967 census, Tanzania has not been collecting direct information on languages. In Column 6, we report the share of Bantu language speakers from Molnos (1969: 48), and the share of non-Bantu language speakers from the 1957 census. Unfortunately, shares are not provided. The reported values are understood as shares of the population in the 1967 census (12313469 people).

[3]According to Leclerc (2010) and Githiora (2008), 95% of the population speaks Swahili. Adegbija (1994, p. 7) reports that 0.6% and 90% speak Swahili as first and second language, respectively. According to Abdulaziz (1970), 10% of the population speak Swahili as their mother tongue and about 90% are bilingual in Swahili and a vernacular language. The share of the population that is literate in Swahili is 69.8% in the 2002 population census (National Bureau of Statistics 2006) and 70.8% according to the 2012 population census (National Bureau of Statistics 2014). However, percentages are understood as the share of the population aged 10 or above in the former and as the share of the population aged 5 or above in the latter. Reading and writing abilities are not tested.

[4]The share of population speaking English is 4.5% in Leclerc (2010), 10% in Graddol (1997, p. 11), and 11% in Crystal (2004, p. 64.). The share of the population which have any knowledge of English is 15% in Abdulaziz-Mkilifi (1972). The share of the population literate in English is 10.8% in the 2002 census (National Bureau of Statistics 2006) and 14.2% in the 2012 census (National Bureau of Statistics 2014).

Uganda

Official languages: English, Swahili
Ethnic diversity: 0.896
Linguistic diversity: 0.896
Index of Communication Potential: 0.484
The number of languages in Ethnologue: 41

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as ethnicity (census 2002)[1]
Ganda[2]	21.7	21.65	27.3	48.95	14.27 (3.45 as L2)	17.3
Soga	13.34	13.34	2.14	15.48	7.12	8.6
Nyankole	11.1	11.1	5.19	16.29	8.05	9.8
Lango	9.54	9.54	0.42	9.96	5.15	6.2
Acholi	6.71	6.75	0.79	7.54	4.04	4.8
Teso	5.29	5.29	0.49	5.78	5.4	6.6
Chiga	4.7	4.7	2.4	7.1	5.46	7
Nyoro	4.75	4.75	2.86	7.61	2.3	-
Masaaba	4.57	4.57	0.6	5.17	3.87	-
Tooro	3.25	3.25	5.8	9.05	1.69	-
Lugbara	2.81	3.09	1.38	4.47	2.75	4.3
Madi	2.7	2.7	0.4	3.1	1.02	-
Alur	2.35	2.35	0.46	2.81	2.13	-
Konzo	2.25	2.25	0.46	2.71	2.1	-
Swahili[3]	-	0.02	15.35	15.37	0.0001	-
English[4]	-	-	51.17	51.17	no data on L1 speakers (8.64 as L2)	-

General notes: Unfortunately, reports on the 2002 population census and the Demographic and Health Surveys cannot be utilised to derive information on literacy rates per language. However, Ladefoged (1971) provide some information on language proficiency in Swahili, Luganda and English based on a large-scale sociolinguistic survey series conducted in several East African countries (Ethiopia, Kenya, Uganda, Tanzania, and Zambia). The study in Ladefoged (1971) is based on more than 2000 personal interviews from all over the country. Since the sample was not fully balanced for factors as age, sex, or tribe, the results must be considered as tentative approximations.

[1]Source: Uganda Bureau of Statistics (2006).

[2]The share of the population speaking Ganda is 60% (30% as L1 and 30% as L2) in Baker and Jones (1998, p. 368), 38% (18% as L1 and 20% as L2) in Adegbija (1994, p. 7), and 22% (17% as L1 and 5% as L2) in the previous edition of the Ethnologue (Lewis 2009). In Ladefoged (1971, pp, 24-25) the share of the respondents that is able to communicate in Ganda (as L1 or L2) was 39%. According to Namyalo (2010)'s estimates, the number of people speaking Ganda either as native or second is about 15 million, which is about 50% of the country population in the Ethnologue (Lewis et al. 2014.).

[3]The share of the population speaking Swahili is approximately 42% (<1% as L1) in Albaugh (2014, p. 277). The reported share is 35% in Adegbija (1994, p. 7), 90% in Leclerc (2010), and 20% in Baker and Jones (1998). The share of the respondents that is able to communicate in Swahili (as L1 or L2) is 35% in Ladefoged (1971 p. 24-25).

[4]The share of the population speaking English 6% in Leclerc (2010), 9% in Graddol (1997, p. 11), and 10% in Crystal (2003, p. 65). In Ladefoged (1971, pp. 24-25) the share of the respondents that is able to communicate in English (as L1 or L2) was 21%.

Zambia

Official language: English
Ethnic diversity: 0.884
Linguistic diversity: 0.872
Index of Communication Potential: 0.663
The number of languages in Ethnologue: 46

(1) language/ ethnicity (AB)	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) share as L1 (Ethnologue)	(7) share as ethnicity (census 2010)[1]	(8) share as L1 (census 2010)[1]	(9) share as L2 (census 2000)[2]	(10) share as L1 in Williams (1998)[3]	(11) share as L2 in Williams (1998)[3]
Bemba[3]	26.4	28.91	34.95	63.86	27.36	20.99	33.38	20.2	30.8	25.4
Tonga	15.32	14.99	8.37	23.36	9.55	13.56	11.38	4.4	16.1	7.1
Chewa/ Nyanja[4]	7.41	10.39	38.33	48.72	15.66	7.83	18.29	21.8	16	26.1
Lozi	8.69	8.33	6.38	14.71	4.38	5.72	5.48	5.2	9.3	7.9
Tumbuka	4.79	4.66	3.4	8.06	2.63	4.4	2.55	1.3	-	-
Nsenga	4.69	4.42	4.33	8,74	2.91	5.28	2.94	1.6	-	-
Lunda	3.40	3.37	2.76	6.13	1.54	3.54	2.05	1.3	2.9	2.4
Ngoni	4.21	3.45	-	-	-	4.02	0.7	1.2	-	-
Kaonde	2.94	2.87	5.6	8.47	1.48	2.93	1.85	1.8	3.4	3.7
Luvale	2.73	2.44	4.69	7.13	1.23	2.18	1.53	1.9	5.9	2.2
English[5]	0.09	0.09	40.44	40.53	0.8 (12.93)	0.0002	1.65	26.3	-	-

General notes: Information on the dimensions of the language situation is provided in Marten and Kula (2008). Although Ngoni is a separate ethnic and linguistic group in Afrobarometer and the 2010 population census, it is most commonly understood as a large cultural group or a tribe which members speak various languages. In Ethnologue (Lewis et al. 2014), Ngoni can refer to a dialect or Chewa, Tumbuka and Nsenga. Additional information on Zambian languages is found in Ohanessian and Kashoki (1978). Unfortunately, the 2010 population census and the Demographic and Health Survey reports are not applicable to derive information on literacy per language.

[1]Central Statistical Office of Zambia (2012)

[2]Marten and Kula (2008)

[3]Williams (1998)

[4]According to Adegbija (1994, p. 10), Bemba is spoken by 31% and 25% as L1 and L2, respectively. According to Baker and Jones (1998, p. 369) the share of people speaking Bemba is 60%.

[5]The share of Chewa/Nyanja speakers is 53% (11% as L1 and 42% as L2) in Adegbija (1994, p. 10).

[6]The share of the population speaking English is more than 30% in Baker and Jones (1998, p. 369), 11% in Graddol (1997, p. 11), and 19.5% in Crystal (2003, p. 65).

Zimbabwe

Official language: English
Ethnic diversity: 0.827 (groups within Shona are distinguished)
Linguistic diversity: 0.331 (dialects of Shona are not treated as separate languages)
Index of Communication Potential: 0.871
The number of languages in Ethnologue: 21

(1) language/ ethnicity	(2) share as ethnicity (AB)	(3) share as home language (AB)	(4) share as additional language (AB)	(5) total share (AB)	(6) as L1 (Ethnologue)
Ndebele	12.22	12.61	13.87	26.48	11.95
Shona[1]	32.97	80.77	10.92	91.69	82.47 (13.87 as L2)
Zezuru	14.6	-	-	-	24.66
Korekore	7.33	-	-	-	13.1
Manyika	5.54	-	-	1.91	6.64
Ndau	5.08	-	-	1.34	6.17
English[2]	-	0.35	43.95	44.3	1.93 (40.85 as L2 speakers)

General notes: Surveys in Zimbabwe (population and housing censuses and the Demographic and Health Surveys) do not collect data on languages. Ethnicity is rather understood as race (African, European, Coloured, and Asian). In the Afrobarometer questionnaire, the subgroups within Shona are listed separately in Q79 on ethnicity, but not listed separately as dialects in Q3 on home languages. Ndau and Manyika are considered as separate, yet partially intelligible languages with Shona in Lewis et al. (2014).

[1]According to Nyika (2008, p. 459) and Adegbija (1994, p. 11) 75% of the population speaks Shona.

[2]According to Graddol (1997, p. 11) 28% of the population speaks English. The corresponding share is 49% in Crystal (2003, p. 65).

4 Languages, communication potential and generalized trust in Sub-Saharan Africa: Evidence based on the Afrobarometer Survey[40]

Abstract

The goal of this study is to investigate whether speaking other than home languages in Sub-Saharan Africa promotes generalized trust. Based on various psychological and economic theories, a simple model is provided to illustrate how languages might shape trust through various channels. Relying on data from the Afrobarometer Project, which provides information on home and additional languages, the Index of Communication Potential (ICP) is introduced to capture the linguistic situation in the 20 sample countries. The ICP, which can be computed at any desired level of aggregation, refers to the probability that an individual can communicate with a randomly selected person in the society based on common languages. The estimated two-level hierarchical models show that, however, individual level communication potential does not seem to impact trust formation, but living in an area with higher average communication potential increases the chance of exhibiting higher trust toward unknown people.

Keywords: Sub-Saharan Africa, generalized trust, communication potential, diversity measurement, multilevel modeling

---

[40] This study is published in Social Science Research, 49(1), 141-155. Online available at http://www.sciencedirect.com/science/article/pii/S0049089X14001604. The author is grateful to the anonymous reviewers at Social Science Research for their helpful and directive comments. The author would also like to acknowledge Peter Foldvari (Utrecht University, the Netherlands) for invaluable advice and help with regard to methodological issues.

4.1 Introduction

Trust as part of the social capital and a measure of social cohesion has been acknowledged to have beneficial effects on several aspects of human life. At the macroeconomic level it is associated with physical and human capital accumulation (Dearmon and Grier 2011, Papagapitos and Riley 2009), the rate of economic growth (Horváth 2013, Zak and Knack 2001), and the magnitude of economic volatility (Sangnier 2013). Generalized or social trust also might promote democracy and democratic stability (Newton 2001, Inglehart 1999, Putnam et al. 1994) and potentially contributes to human development (Özcan and Bjørnskov 2011) and personal happiness (Tokuda et al. 2010).

A recent strand of the literature aims to identify factors that determine trust itself instead of focusing on its positive societal effects. Among the numerous investigated variables, ethnic (racial) and linguistic diversity have been given special attention (Dincer 2011, Tsai et al. 2010, Bjørnskov 2008, Putnam 2007, Delhey and Newton 2005) partly due to the increased importance of immigration from developing to developed countries in the past decades contributing to higher cultural diversity in the host countries (Gerritsen and Lubbers 2010, Hooghe et al. 2009). However, research on the relationship between ethnic and linguistic diversity and trust focusing on less developed and traditionally diverse societies is less available (Huhe 2014). Our study has several contributions. First, by focusing on Sub-Saharan Africa we make it possible to expand our knowledge of factors affecting generalized trust based on evidence from a less developed region characterized by high diversity. Second, we broaden the spectrum of the previously investigated trust-promoting factors by taking languages into account. Although it is not surprising that African people usually master more than one language (Laitin 1992), studies devoted to explaining the underdevelopment of Africa rely on fragmentation measures (Taylor and Hudson 1972, Alesina et al. 2003, Fearon 2003) that fail to take this fact into account. In this chapter the language factor is measured by the Index of Communication Potential (ICP), which, utilizing data provided by the Afrobarometer Project, is able to account for not only home languages but the whole individual language repertoire. Third, we apply multilevel or hierarchical modeling methods to shed light on the possible relationship between languages and trust in Sub-Saharan Africa. Although this technique is widely utilized in health research, political sciences, and sociology, it is less frequently used in economics, even though it has some conceptual and statistical advantages over classical regression analysis. The results from the two-level multilevel models with ordered dependent variable provide evidence that an increasing communication potential indeed promotes trust in Sub-Saharan Africa. While the individual-level coefficients of the communication potential do not turn out to be statistically significant, at a regional level, there is a significant positive effect on trusting behavior.

The chapter proceeds as follows. The next section is devoted to the discussion of the possible channels through which languages might foster generalized trust. Based on available economic and psychology literature, four hypotheses are articulated. Section 3 describes the data and variables utilized in the empirical analysis in Section 4. The last section draws conclusions based on this research.

4.2 Theoretical background

Trust as an abstract concept can be approached from several aspects. Although trust formation and trusting behavior are topics of several disciplines, Rousseau et al. (1998) argue that there is a general consensus that trust refers to expectations about the risk of engaging in an interaction with someone else. The literature distinguishes two basic types of trust depending on the subjects of the trusting behavior. Particularized or personalized trust refers to trust shown toward acquaintances and relatives, while generalized or social trust is the trust toward unknown people (Nannestad 2008, Bjørnskov 2006, Uslaner and Conley 2003). These concepts are referred to as "thick" and "thin" trust by Putnam (2000) and "strategic" and "moralistic" trust by Uslaner (2002), respectively. Another approach distinguishes among the types of trust based on its sources or fundamentals (Jones 1996, McAllister 1995). Cognition- or knowledge-based trust is built upon available information on others' behavior under certain circumstances, while emotion- or affect-based trust originates from one's own emotions and sense of others (Chua et al. 2008). A third research line aims to reveal whether trust is the property of individuals and determined by personal traits or is it rather rooted in the social context or the social system (Newton 2004, Delhey and Newton 2003). The puzzle as to how the aforementioned different components of the same abstract concept relate to each other has inspired many scholars (e.g. Freitag and Bauer 2013, Freitag and Traunmüller 2009, Uslaner 2000-2001). Since generalized trust is related to unknown people, it is logical to assume that knowledge affecting determinants, such information as that derived from repeated interactions and cooperation, play a larger role in building particularized trust. Although it is well established that personalized and social trust are interconnected, results regarding the direction and magnitude of this relationship are mixed. One theory proposes that at a certain point trust built upon personal experience is transferred to unknown individuals or groups (Putnam 2000). On the contrary, another theory derived from network studies (Granovetter 1973) points out that strong inter-group relationships limit the success of outside group interactions.

Language can be seen as a basic tool for interactions that facilitates information flow between individuals, reduces the costs of cooperation, and ultimately fosters repeated interaction and helps to solve collective action problems, which has a positive payoff in the long run (Smith 2010, Ostrom 2000). This process is likely to increase

trust toward known people through building up experience and accumulating information on the behavior of persons that belong to a different group in terms of primary language. Psychology offers various theories on how increased personalized trust may spill over to a higher degree of generalized trust. It is well established that contrary to the rational choice theory, people are subject to cognitive biases when making judgments about situations and other individuals (Kahnemann et al. 1982). When information processing is impractical or the information is imperfect, people tend to use heuristics or mental shortcuts based on previous experiences to ease the decision-making process. Based on the law of small numbers (generalization based on a few examples) (Tversky and Kahnemann 1971) and the outgroup homogeneity hypothesis (the other group is more homogenous than one's own) (Messick and Mackie 1989, Quattrone and Jones 1980), people are likely to have optimistic attitudes toward a group when they have positive experience with some of the group members.

Social psychology has also documented that people are attracted by others who are similar to them in certain aspects (Buss 1985, Byrne 1971, Bond et al. 1968, Newcombs 1961). Moreover, perceived similarity and attractiveness might reduce uncertainty that people attribute to others (Parks and Adelman 1983). As Leeson (2005) suggests, in certain situations language learning can be considered as one's effort to be perceived as more similar by others. However, since language is not an easily observable attribute of individuals and can only be revealed by making personal contact, we assume that the beneficial effects of increased similarity due to language learning works again through the link of personalized trust.

The aforementioned theories derived from microeconomics and predominantly psychology offer us a framework in which to theorize that individual behavior of learning a second language indeed promotes generalized trust via spillovers from personal trust. However, as several previous studies argue, individual trusting behavior is dependent on both personal characteristics and the social environment. Obviously the possible returns in terms of lowered transaction costs as a result of acquiring new languages depend on decisions made by others as well. A rational decision maker, even in the presence of informational imperfections, will first weigh the possible gains from learning new languages against its significant costs. The possible gains are not obvious, however, unless one has information on the language learning strategy of others. This information (expectation regarding the decision of others) can be gained from personal experience. The probability that an individual will encounter people from another language group with whom they have at least one common language depends on the average language repertoire of the population. Based on these considerations we derive two hypotheses. First, language knowledge affects generalized trust positively (Hypothesis 1) and second, due to imperfect information individual generalized trust is more dependent on the expectation that is derived from personal experience (Hypothesis 2).

Acquiring second languages helps to build social cohesion through an additional channel that works more at the macro than the micro level. Although several studies

have shown that social heterogeneity defined in ethnic, linguistic, religious, and other terms, is likely to reduce trust and social capital (Putnam 2007, Alesina and La Ferrara 2002, Rice and Steele 2002), cross-cutting cleavages (groups defined along one dimension overlap defined by another) can moderate the detrimental effect of such fragmentations (Dunning and Harrison 2010, Coser 1956). Since second languages can influence identity (Aspachs-Bracons 2007, 2008), they are a sufficient basis for cross-cutting cleavages (Hypothesis 3). Moreover, second languages facilitate information flow not only among individuals, but also among groups.

Lastly we hypothesize that the relationship between languages and generalized trust might be non-linear (Hypothesis 4). First, language can be seen as a product; its value or return increases as the number of users grow (network externalities) (Church and King 1993). Increasing returns are expected to affect actual and expected language choices, hence generalized trust. Treating languages as a network good also implies that there is a minimum number of speakers (critical mass) below which network externalities are not present (Economides and Himmelberg 1995). Secondly, based on the cross-cutting cleavages theory, we assume that the severity or the depth of social fragmentation along a certain dimension might weaken the binding power of second languages.

4.3 Data and methodology

4.3.1 The Afrobarometer and the multilevel method

The data used in this study are from the Afrobarometer Survey Project[41], which has increasingly been utilized in political, development and even historical studies related to Africa (Eifert et al. 2010, Nunn 2010) in past years. Although the primary goal of this non-partisan survey is to map the political atmosphere and attitudes toward democracy, it also contains general socio-demographic information of the respondents (and interviewers) that makes this source valuable for more general social research. The sample is designed as a representative cross-section of all citizens of voting age in a given country. Since it is only Round 4 that provides information on second languages, this study must be limited to that. Countries included in the dataset together with the number of observed units at the level of individuals and geographical regions are presented in Appendix 4A. Since the questionnaire for Cape Verde is different from those of other countries in relevant aspects, it has been excluded from the final analysis.[42]

[41] www.afrobarometer.org.

[42] Q79 utilized to compute ethnic fragmentation is different in Cape Verde. It is not related to ethnic membership per se, but respondents are asked to select the most important dimesion of their self-identification. Among the several possibilities (gender, occupation, geographical area, etc.) "race" is one option.

Instead of conducting linear regression analysis, we make use of the theoretical and statistical benefits of the multilevel modeling technique that can be considered as a basic analytic tool in health research (see for instance Kim and Kawachi 2006, Diez-Roux 2000, O'Campo et al. 1997). This approach has recently begun to gain popularity in social and political sciences (see for instance Finseraas 2008, Chung and Muntaner 2007, Fieldhouse et al. 2007, Weldon 2006, Ulmer and Johnson 2006). Huhe (2014 p. 582-584) lists several reasons why trust research would benefit from using multilevel modeling. As Delhey and Newton (2003) highlight, one strand of the literature stresses the importance of personal characteristics in determining individual trust, while another argues that context-related factors play a more influential role. Multilevel modeling enables the testing and reconciling of these competing views by incorporating individual and contextual determinants in a single analytical framework. In addition, the multilevel method can effectively incorporate abstract multilevel concepts that are quite common in social sciences. For instance income inequality might be captured and elucidated at the level of individuals by measuring perceived inequality on the one hand and at the contextual level by computing the Gini coefficient on the other. Moreover, various phenomena (e.g. religion) are documented to impact attitudes through different channels at the personal (upbringing, values) and the contextual (religious cleavages, transmitting norms to the broader society) level. In the theoretical framework set up in the previous section it was argued that this last point applies to languages as well. And finally, since it is multilevel in nature, the hierarchical modeling technique helps to avoid the trap of statistical fallacies (ecological, individualistic), i.e. conducting research at an analytical level and drawing conclusions in relation to another (Kramer 1983, Robinson 1950). The multilevel approach also increases the reliability of the estimation through handling observation dependence caused by the common contextual factors. Further technical details and methodological advantages are provided in Diez-Roux (2000), Snijders and Bosker (1999) and Hox (1995).

Two-level hierarchical models are estimated in this study; individuals are classified as level-one, and geographical areas are considered as level-two units. The "region" variable of the Afrobarometer corresponds with the highest official sub-national level administrative units and the geographical partitioning applied in the latest national censuses at the time the survey was conducted. Although most African countries have progressed toward more decentralized governance systems since the mid-1980s, the pace and the form of implementation show a very uneven picture across countries (USAID 2010). The most populous countries (South Africa, Nigeria) tend to be federal with political decentralization at multiple levels and considerable fiscal devolution. Countries formerly colonized by France (e.g. Burkina Faso, Mali) are still relatively centralized with limited fiscal and administrative authority. Using regions as the second level of our analysis makes it possible to control for diversities within country borders. Regions are small enough to exhibit a large degree of heterogeneity in terms

of languages and socio-economic characteristics, while large enough so that the majority of transactions take place within their boundaries.

Finally, according to the rule of thumb proposed by Hooghe et al. (2009 p. 207) based on Maas and Hox (2005), a proper multilevel analysis requires at least 30 observations at each level. Choosing a higher or lower level administrative unit as second level would result in a less sufficient sample: the number of countries is only 19, and the majority of the lower geographical units (variable "district" in the Afrobarometer) contain fewer than 30 respondents.

4.3.2 The dependent variable: generalized trust

To construct the dependent variable for the empirical investigation, Question 84C was utilized (Q84C: How much do you trust other Beninese/Ghanaian etc?) to proxy the level of generalized trust in Sub-Saharan Africa. The respondents could select an answer on a four point scale ((0) Not at all, (1) Just a little, (2) I trust them somewhat, (3) I trust them a lot) with the additional possibility of choosing "Don't know" and "Refused to answer." The analysis was restricted to answers that could be measured on an ordinal scale, and the remaining possibilities were considered as missing values. This strategy enabled the estimation of ordered logit models. Table 4.1 displays the distribution of the above mentioned possible answers among respondents and the ranking of countries based on the mean trust values.

The question regarding generalized or social trust in the Afrobarometer Survey is similar but not identical to the most widely exploited question provided by the World Value Survey, that is "Generally speaking, would you say that most people can be trusted or you can't be too careful in dealing with other people?". The possible answers (yes or no) are eligible for an analysis as a binary dependent variable. The question in this form is quite vague and does not include who is meant by "most people." Under such circumstances, it might be considered as a valid and reliable measure of generalized trust. (For a good summary of the available literature on this issue, see Bjørnskov 2006).

Table 4.1 Generalized trust in twenty Sub-Saharan African countries

country	(0) Not at all	(1) Just a little	(2) I trust them somewhat	(3) I trust them a lot	mean (individual standard deviation)	ranking based on the mean
Benin [1189]	46.44	32.21	13.18	8.17	0.831 (0.946)	20.
Botswana [1199]	26.05	39.36	22.92	11.67	1.202 (0.957)	11.
Burkina Faso [1149]	20.24	29.60	20.48	29.68	1.596 (1.114)	6.
Cape Verde [1250]	42.20	27.51	23.93	6.36	0.944 (0.956)	19.
Ghana [1175]	15.61	30.73	29.28	24.38	1.624 (1.017)	5.
Kenya [1082]	12.77	46.45	28.07	12.71	1.407 (0.867)	10.
Lesotho [1193]	22.75	31.27	22.27	23.71	1.469 (1.086)	8.
Liberia [1194]	27.05	44.99	18.74	9.22	1.101 (0.904)	16.
Madagascar [1240]	27.94	46.95	18.45	6.66	1.038 (0.854)	17.
Malawi [1186]	15.32	28.88	22.41	33.40	1.739 (1.081)	2.
Mali [1214]	13.70	27.50	32.89	25.91	1.71 (0.999)	3.
Mozambique [1180]	36.90	30.17	12.33	20.61	1.167 (1.136)	14.
Namibia [1184]	30.73	32.10	25.18	11.98	1.184 (1.003)	12.
Nigeria [2284]	36.08	33.59	24.34	5.98	1.002 (0.919)	18.
Senegal [1151]	18.13	24.21	26.52	31.14	1.707 (1.093)	4.
South Africa [2372]	24.75	40.26	27.90	7.09	1.173 (0.883)	13.
Tanzania [1092]	4.41	18.37	53.64	23.59	1.964 (0.771)	1.
Uganda [2424]	26.71	42.74	20.39	10.16	1.14 (0.926)	15.
Zambia [1185]	22.77	32.49	24.22	20.53	1.425 (1.054)	9.
Zimbabwe [1192]	16.98	27.76	36.88	18.38	1.567 (0.976)	7.
Total [27135]	24.9	33.62	25.48	15.99	1.324 (1.015)	-

Note: The share of respondents in each answer category is in percentages, square brackets contain the number of respondents per country. In column 5, individual standard deviation from the country mean is presented in parentheses.

4.3.3 The main explanatory variable: Index of Communication Potential (ICP)

Based on Q3 (What is your home language?) and Q88E (What languages do you speak well?) of the fourth round of the Afrobarometer Project, the Index of Communication

Potential (ICP)[43] has been constructed. Accounting for multilingualism, this indicator is expected to provide a more realistic picture of the linguistic situation in Sub-Saharan Africa compared to indices that rely on home languages only.

The beneficial property of this measure, i.e., it can be computed at the level of individuals and any desired higher aggregation levels, enables for the testing of hypotheses outlined in the previous section. Technical details on the construction of the Index of Communication Potential are presented in Appendix 4B. The individual communication potential equals the probability that a respondent can communicate with another randomly selected person within the country based on both common home and second languages (Eq. 4B.1 and Eq. 4B.2 of Appendix 4B). The individual communication potential measures and their regional means (Eq. 4B.3 and Eq. 4B.4 of Appendix 4B) are used as explanatory variables in the first and second level equations of our estimated hierarchical models. Table 4.2 shows a summary on the above mentioned variables. The aim of the first column of Table 4.2, which presents the most widely spoken languages, is to show which languages contribute to the largest extent to the country's ICP.

In relation to the theoretical framework, individual level ICP can be considered as the actual share of cohabitants with whom an individual can communicate within the society, but due to the presence of incomplete information, individuals are not aware of its precise value. However, it is assumed, based on their personal experiences, that people have expectations about the average linguistic repertoire, which can be proxied by the regional average ICP. Furthermore, the regional ICP can be understood as an alternative to existing diversity measures that focuses more on the bridging effect of common languages than diversity along the traditional dimensions of ethnicity, home language, etc.

For clarification it is important to state that in the empirical analysis regional level communication potential indices do not refer to the probability that two randomly selected individuals of the same administrative region are able to communicate with each other but show the simple regional means of individual communication potentials discussed above. Thus, for instance, the value of 0.368 in Alibori (Benin) in Table 4.2 might be interpreted that a representative person in that region is expected to be able to communicate with about 37% of the total Beninese society.[44]

[43] While in the case of Q3, respondents are required to choose from a pre-defined list of the most widely spoken home languages, Q88E is based on self-reporting.

[44] We have computed an additional type of communication potential both at the level of individuals and regions to show the communication possibilities within geographical regions instead of countries. However, since variables calculated this way show a high mean (over 0.9) and relatively low standard deviation (0.125 and 0.173 respectively), they are less efficient for empirical investigation.

Table 4.2 The Index of Communication Potential in Sub-Saharan Africa

country	largest language (first and second speakers altogether)	individual level	regional level			ranking based on individual ICP
		mean (individual standard deviation)	mean (regional standard deviation)	minimum ICP [region]	maximum ICP [region]	
Benin	Fon 59.1%	0.573 (0.216)	0.580 (0.127)	0.368 [Alibori]	0.737 [Littoral]	16.
Botswana	Setswana 99.2%	0.982 (0.087)	0.984 (0.015)	0.932 [Chobe]	0.993 [there are 5 regions with the highest ICP]	3.
Burkina Faso	Moore 67%	0.602 (0.238)	0.602 (0.121)	0.331 [Sahel]	0.741 [Central]	15.
Ghana	Akan 55.7	0.621 (0.23)	0.615 (0.147)	0.299 [Northern]	0.765 [Eastern]	12.
Kenya	Swahili 94.4%	0.907 (0.168)	0.917 (0.049)	0.759 [North-Eastern]	0.964 [Nairobi]	4.
Lesotho	Sesotho 100%	1.000 (0.000)	1.000 (0.000)	not applicable		1.
Liberia	English 72.4%	0.609 (0.285)	0.598 (0.128)	0.281 [Lofa]	0.768 [Margibi]	14.
Madagascar	Malagasy 100%	1.000 (0.000)	1.000 (0.000)	not applicable		1.
Malawi	Chewa 92.9%	0.877 (0.195)	0.841 (0.138)	0.670 [Northern]	0.926 [Central]	6.
Mali	Bambara/ Bamanankan 87.3%	0.786 (0.233)	0.803 (0.144)	0.119 [Kidal]	0.884 [Bamako]	8.
Mozambique	Portuguese 78.7%	0.700 (0.260)	0.697 (0.080)	0.467 [Gaza]	0.794 [Maputo City]	9.
Namibia	English 76.3%	0.815 (0.164)	0.816 (0.060)	0.716 [Kunene]	0.911 [Omusati]	7.
Nigeria	English 64.5%	0.626 (0.215)	0.622 (0.078)	0.397 [Ondo]	0.803 [FCT]	11.
Senegal	Wolof 92.3%	0.888 (0.179)	0.892 (0.078)	0.620 [Matam]	0.937 [Dakar]	5.
South Africa	English 65.6%	0.612 (0.265)	0.606 (0.127)	0.381 [Limpopo]	0.746 [Gauteng]	13.
Tanzania	Swahili 99.4%	0.990 (0.064)	0.990 (0.014)	0.927 [Manyara]	0.997 [Mwanza]	2.
Uganda	Luganda 48.9%	0.489 (0.236)	0.468 (0.113)	0.316 [North]	0.589 [Central]	17.
Zambia	Bemba 63.8%	0.643 (0.232)	0.643 (0.120)	0.462 [Eastern]	0.765 [Lusaka]	10.
Zimbabwe	Shona 90.1%	0.870 (0.182)	0.871 (0.133)	0.517 [Matebeleland North]	0.946 [Midlands]	7.

Note: Parentheses show standard deviations (in column 1 at the level of individuals, in column 2 at the level of regions). Square brackets contain the names of regions with the minimum and maximum level of ICP within each country.

4.3.4 Individual and regional covariates

Before proceeding with the detailed introduction of variables utilized in the empirical analysis, a few related remarks should be made in advance. Since the literature lacks a general theory on trust formation, most studies implement exploratory statistical analysis to gain more insight on possible trust-enhancing determinants. Although a huge number of variables have been tested, their exact role in shaping trust has still not been completely established. Moreover, empirical findings and the sign and significance of the estimated coefficients vary by methodology and the characteristics of the society under examination. In this section we account for the most widely acknowledged potential determinants of trust as presented by Huhe (2014), Hooghe et al. (2009), Bjornskov (2008) and (2006), Delhey and Newton (2005). Variables are summarized in Table 4.3.

Basic sociodemographic factors are seen to play a crucial role in trust formation; nevertheless, their impact is not straightforward. While older people are predominantly found to be more trusting (Traunmüller 2011, Hooghe et al. 2009, Paxton 2007), the effect of gender is quite ambiguous. (Mewes (2014) provides an extensive overview on findings related to various gender issues and trust). Religion will be discussed later.

Economic development and additional characteristics of a modern society are suggested to positively influence attitudes toward others. Banfield (1958, p.110 referred to in Delhey and Newton 2005) argues that the wealthier the society the more able the people are to take the risks connected to trusting behavior and the less necessary and rewarding the untrustworthy behavior is. Empirical investigations support this relationship (Wang and Gordon 2011, Alesina and La Ferrara 2002, Zak and Knack 2001, Knack and Keefer 1997). Widespread education as an additional consequence of the modernization process is expected to increase trust through its socializing effects as well as making people more informed and capable of information processing (Wang and Gordon 2011, Knack and Keefer 1997). Although urbanization is part of the modernization process, its positive effect is hardly documented (Alesina and La Ferrara 2002:221, Knack and Keefer 1997:1283). In fact, smaller communities tend to exhibit higher levels of trust (Putnam 2000:205, House and Wolf 1978).

Q1, Q101 and "urbrur" (urban or rural) variables of the Afrobarometer provide information on the age, gender and location of the residence of respondents, respectively. The originally categorical variable (Q89) referring to the educational level (no formal schooling, informal schooling only, some primary schooling, primary schooling completed, some secondary school/high school, secondary or high school completed, post-secondary qualifications other than university, some university, university completed, post-graduate) is transformed to gain a continuous variable

measured in years.[45] In this way unnecessary dummy variables, which reduce the degrees of freedom of the estimation, can be avoided.

Since the Afrobarometer does not include questions on income or wealth, we use the concept of the Lived Poverty Index (LPI) (Mattes 2008) to gain some kind of socioeconomic status measure that substitutes for the standard GDP per capita in the analysis. Information on how many times the respondents had to struggle with the lack of basic necessities (such as food, water, medical care, cooking fuel and cash income) is derived from Q8A to Q8E. Factor analysis is applied to extract the first factor from responses to the five items that account for 53.34 percent of the common variance across all items. As a result of the ordering of the possible answers, higher values refer to more frequent struggling in everyday life. In order to gain easily interpretable values, the variables are standardized on the whole sample to achieve a zero mean and unit standard deviation. Therefore, negative signs are considered as a label of higher well-being compared to the average individual. For more details see Appendix 4C. The regional means are included in the regional level equation to control for the possible contextual effects of a relatively wealthy society. According to this indicator, North Pemba (-1.005) in Tanzania and Matam (1.324) in Senegal seem to be the wealthiest and the poorest regions, respectively (Table 4.3).

Psychological and social studies have shown that attitudes and behavior, thus trust itself, are determined by not only actual circumstances of individuals but also by personal beliefs and subjective factors (Huhe 2014, Bardi and Schwartz 2003). Therefore, in parallel to the socioeconomic status variable that is supposed to capture the actual living circumstances, the life satisfaction of individuals has been incorporated. Utilizing Q4B (In general, how would you describe your present living conditions?) a binary variable is created to measure how respondents evaluate their living conditions. Answers referring to very bad and fairly bad situations have been recoded into one, while the remaining answers (neither good nor bad, fairly good, very good) have been recoded to zero. Combining findings of the psychological literature on the importance of subjective factors with the modernization theory, we expect that poorly evaluated living conditions lead to reduced social trust.

[45] We used the Foreign Credit Class Base Education Database (http://www.classbase.com/) and the websites of universities to gain information on the duration of primary, secondary and tertiary education in each country. Three main challenges occurred. First, we had to decide how to assign years of education to the started but not finished categories; second, how to understand secondary education; and third, how to incorporate vocational training into the system. Unfinished categories take the average value of the completed categories above and below. Most African countries maintain a divided system of secondary education with a lower or junior and an upper or senior level. Consequently, when an individual reports completion of secondary education, it is not straightforward if he or she has completed only the junior level or both. Hence, we use the average of the years required to finish the junior and the senior level as duration of secondary schooling. The other problem is caused by the diverse system of vocational training. Certain types of vocational training require accomplished senior level schooling, while others do not. Thus, post-secondary education (that means mainly vocational training) gets the same value as the senior level secondary schooling. Postgraduate education is understood as one additional year after completing master studies.

The importance of voluntary clubs and associations in trust building was previously argued by Tocqueville and Mill in the nineteenth century (Putnam 2000). While some document that members of voluntary organizations are indeed more trusting (Stolle and Rochon 1998, Brehm and Rahn 1997), others argue that the relationship is weak or insignificant (Delhey and Newton 2005, Freitag 2003). Moreover, Paxton (2007) stresses that the type and characteristics of associations also matter. She finds that while connected associations promote trust both at the individual and country level, the effect of isolated associations is weaker at the individual and negative at the country level. Q22B of the Afrobarometer provides information on membership in a voluntary association or community group (other than religious). Answers referring to inactive or active membership and official leadership are recoded into one; non-membership is recoded into zero. Although the Afrobarometer does not explain what is exactly meant by inactive membership, it is assumed that this status translates into more social interaction compared to non-membership. Following Paxton (2007), association membership at both individual and regional levels have been included.

It is recognized that religion affects trust through various channels. In addition to its direct influence on individuals as an important sociodemographic characteristic, religion can influence the context in which social trust formation takes place by transmitting certain values and norms on the one hand, and providing basis for cleavages on the other (Traunmüller 2011, Delhey and Newton 2005). The wide range of more than 30 religions that respondents in the Afrobarometer might choose in Q90 was collapsed into four main categories (non-religious, Christian, Muslim, and other) in the empirical analysis. Although theoretical considerations and empirical findings suggest that values spread by Protestantism are more trust promoting compared to other religions (Fukuyama 1995, 2000), we do not distinguish between them. The main reason for this is that 21 percent of respondents reporting Christianity select "only Christian" and do not specify the denomination. The religious variable has been included in three forms in the empirical models. First, as an individual level sociodemographic variable, a dummy variable is used that assumes the value of one if the respondent is Christian. Second, the share of Christians in each region is applied to capture the spread of their religious values. Third, the measure of religious diversity (discussed below) captures the probability that two randomly selected people in the region belong to different religious groups.

Empirical findings related to social heterogeneity and trust might be organized around two main theoretical directions. On the one hand, the conflict hypothesis argues that the presence of ethnic or racial heterogeneity can cause feelings of threat and negative attitudes toward those who are different. The literature provides empirical findings to support this argument (Putnam 2007, Alesina and La Ferrara 2002, Rice and Steele 2002). On the other hand, the contact hypothesis highlights the historical time length that the different communities have been in contact and the frequency of social interactions among the members of different groups, and this can

seriously affect the previous results on harmful effects of heterogeneity (Pettigrew 1998, Stolle and Harell 2013, Stolle et al. 2008). A more detailed introduction into the contact and conflict theory is provided in Gundelach (2014).

The fragmentation indices are based on the Herfindahl formula of concentration that is widely accepted in diversity measurement to capture the probability that two randomly selected individuals in the region belong to different ethnic, linguistic, religious and political groups. Q3, Q79, Q90 and Q86 are utilized to reveal the linguistic, ethnic, religious and political affiliation of individuals.

The Afrobarometer Survey allows us to measure linguistic fragmentation separately from ethnic fragmentation. In diversity measurement it is a common practice to treat linguistic and ethnic group membership as close analogs. Language data are used to proxy ethnic membership when information on the latter is not available (Cheeseman and Ford 2007). The Afrobarometer shows that in certain countries ethnic and linguistic fragmentation might differ significantly; the linear correlation coefficient between them is 0.594. The most remarkable gaps between these two measures are found in Botswana, Lesotho, Madagascar, and Zimbabwe where the Tswana, Sotho, Malagasy and Shona languages are reported as the home languages by many respondents belonging to different ethnic groups. Although less striking, Mali and Senegal show considerable differences in terms of the two measures. Our data do not support the negative expected relationship between the ethnic and linguistic fragmentation and the ICP. The ICP correlates positively but weakly (0.262) with the ethnic diversity measure and the relationship with the linguistic heterogeneity indicator (-0.101) is insignificant at the level of regions. Since the logically presumed high correlations among the three aforementioned indicators are not present, all three measures in the same empirical models have been included.

Income inequality as an additional type of social diversity is robustly found to decrease trust in cross-national studies (Bjornskov 2006, Knack and Keefer 1997). Going beyond the national level, Alesina and La Ferrara (2002) and Gustavsson and Jordahl (2008) reinforce this negative relationship by studying US localities and Swedish counties, respectively. Since standard income inequality (the Gini and percentile based indices) measures cannot be derived from the Afrobarometer, the average individual standard deviation of the socioeconomic status variable from the regional mean is applied as a feasible substitute.

We include the logarithm of the population density in our analysis to proxy the frequency that people belonging to different ethnic, linguistic, religious and political groups meet within a region. Psychological literature has empirically confirmed that if people are repeatedly exposed to a particular stimulus object (in this case the members of different ethnic, linguistic etc. groups), they might show increased preferences for that object (Zajonc 2001). Regional population density data are retrieved from the latest available national censuses at the time that the Afrobarometer Survey was conducted.

And finally, we test whether corruption harmfully affects social trust as proposed by theoretical considerations and empirical evidence (Rothstein and Eek 2009, Delhey and Newton 2005). Relying on variables Q50A to Q50H that reveal the opinion of respondents on the level of corruption among the different players of the public sphere (e.g. the president and his office, local government councelors, police, traditional leaders etc.), a measure similar to the Corruption Perception Index developed by the Transparency International can be created. Although factor analysis in this case could be an appropriate approach, it would result in a high number of missing values due to the relatively low response ratio. Therefore, since the Cronbach's alpha suggests a high internal consistency of these variables and the number of missing answers is the lowest in the case of Q50E (on police), the latter is utilized to capture perceived corruption. Answers indicating that most or all the police are corrupt are recoded into one. The corruption perception is included at both levels. The regional level corruption perception is understood as the share of respondents who believe police to be corrupt.

Table 4.3 Summary of the individual and regional level covariates

covariate	name in regressions	corresponding AB variable	number of observations	mean	standard devation	min	max
Individual-level variables							
age (year)	age	Q1	26116	36.26	14.4	18	110
gender (male=0, female=1)	female	Q101	26449	0.5	0.5	0	1
rural (urban=0, rural=1)	rural	URBRUR	26449	0.645	0.478	0	1
christian (Christian=1)	christ	Q90	26261	0.658	0.475	0	1
education (years)	educ	Q89	26408	6.9	4.66	0	21
socioeconomic status (standardized)	ecstatus	Q8A to Q8E	26059	0	1	-1.374	2.87
life satisfaction (bad=1)	lifesat	Q4B	26308	0.501	0.5	0	1
member in voluntary association (yes=1)	member	Q22B	26221	0.395	0.489	0	1
perceived corruption of police (yes=1)	corr	Q50E	26248	0.147	0.354	0	1
Index of Communication Potential	icp	Q3 and Q88E	26447	0.743	0.264	0.001	1
Regional-level variables							
ethnic fragmentation	reg_ethnic	Q79	255	0.597	0.236	0	0.986
linguistic fragmentation	reg_ling	Q3	255	0.466	0.265	0	0.956
religious fragmentation	reg_relig	Q90	255	0.295	0.195	0	0.685
political fragmentation	reg_politic	Q86	255	0.448	0.197	0	0.832
inequality (standard deviation of socioeconomic status variable)	inequality	Q8A to Q8E	255	0.883	0.154	0.396	1.47
mean economic status	reg_ecstat	Q8A to Q8E	255	0.002	0.45	-1.005	1.324
ln_population density	ln_pop	national censuses	255	4.228	1.6	-1.262	9.038
perceived corruption	reg_corr	Q50E	255	0.506	0.184	0.03	0.96
voluntary membership	reg_member	Q22B	255	0.4	0.163	0.042	0.842
share of Christians	reg_christ	Q90	255	0.655	0.309	0	1
regional ICP	reg_icp	Q3 and Q88E	255	0.737	0.198	0.119	1

Note: All the data except for the population density are derived from the Afrobarometer (AB) Survey.

4.4 Empirical analysis

4.4.1 The model

Although individual generalized trust can be observed as a categorical variable derived from Q84C ($trust_{ir}$), it is assumed that trust is a continuous variable that cannot be observed ($trust^*_{ir}$). The unobserved or latent continuous trust variable is modeled in Eq. 4.1.

$$trust^*_{ir} = \sum_{k=1}^{11} \beta^k x^k_{ir} + \sum_{l=1}^{L} \gamma^l z^l_r + \upsilon_r + \varepsilon_{ir} \,, \qquad \text{(Eq. 4.1)}$$

where x^k_{ir} refers to the *k*th (k=1 to 11) individual-level variable for individual *i* in region *r* and z^l_r refer to the *l*th (*l*=1 to *L*, *L*=11 in Model 1, *L*=12 in Model 2 and *L*=16 in Model 3) regional level variable in region *r* as discussed in the previous section and summarized in Table 4.3. υ_r is a random intercept for region *r* with distribution $\upsilon_r \sim N(0, \tau^2)$, and ε_{ir} is the individual error term with distribution $\varepsilon_{ir} \sim N(0, \sigma^2)$.
The threshold model that determines the observed response in Q84C is introduced in Eq. 4.2.

$$trust_{ir} = \begin{cases} 0 \text{ (not at all) if } trust^*_{ir} \le k_1 \\ 1 \text{ (just a little) if } k_1 < trust^*_{ir} \le k_2 \\ 2 \text{ (I trust them somewhat) if } k_2 < trust^*_{ir} \le k_3 \\ 3 \text{ (I trust them a lot) if } trust^*_{ir} > k_3 \end{cases} \qquad \text{(Eq. 4.2)}$$

We estimate the proportional log odds (ordered logit) for individual *i* in region *r* as in Eq. 4.3.

$$\log\left(\frac{P(trust_{ir} > m)}{1 - P(trust_{ir} > m)}\right) = \sum_{k=1}^{11} \beta^k x^k_{ir} + \sum_{l=1}^{L} \gamma^l z^l_r + \upsilon_r - k_{m+1} \,, \qquad \text{(Eq. 4.3)}$$

where *m* refers to trust categories in Q84C (*m*= 0 to 2) and Eq. 4.2. The model is estimated in Stata 11 using the *gllamm* program. Additional technical details on the multilevel models with ordered dependent variable is provided in Rabe-Hesketh (2004). As proposed by Hooghe et al. (2009 p. 207), regions with fewer than 30 observations (39 out of 255) are excluded from the final analysis.

4.4.2 Results and discussion

Results of estimating Eq. 4.3 are presented in Table 4.4. Model 1 indicates that older people and men show more trust toward unknown people in Sub-Saharan Africa. Christians do not seem to have different trusting behavior compared to those affiliated

with other religions; however, in geographical areas with higher share of Chrisitian population, lower trust categories are more likely.

Table 4.4 Results of the multilevel analysis

	Model 1	Model 2	Model 3
age	**0.009*** (0.001)**	**0.009*** (0.001)**	**0.009*** (0..001)**
female	**-0.155*** (0.026)**	**-0.155*** (0.026)**	**-0.155*** (0.026)**
rural	**0.087*** (0.033)**	**0.087*** (0.033)**	**0.088*** (0.033)**
christ	-0.009 (0.035)	-0.009 (0.035)	-0.009 (0.035)
educ	**-0.014*** (0.004)**	**-0.014*** (0.004)**	**-0.014*** (0.004)**
ecstatus	0.004 (0.015)	0.004 (0.015)	0.004 (0.015)
lifesat	**-0.117*** (0.027)**	**-0.117*** (0.027)**	**-0.118*** (0.027)**
member	**0.241*** (0.027)**	**0.241*** (0.027)**	**0.241*** (0.027)**
corr	**-0.204*** (0.027)**	**-0.204*** (0.027)**	**-0.204*** (0.027)**
icp	0.018 (0.076)	0.017 (0.076)	0.015 (0.076)
reg_ethnic	**0.619*** (0.234)**	**0.536** (0.24)**	**1.487*** (0.393)**
reg_ling	**0.541** (0.216)**	**0.619*** (0.222)**	-0.454 (0.396)
reg_relig	**-0.554** (0.25)**	**-0.582** (0.25)**	**-0.557** (0.247)**
reg_politic	**-1.035*** (0.233)**	**-0.997*** (0.233)**	**-0.837*** (0.238)**
inequality	-0.07 (0.287)	-0.044 (0.287)	-0.087 (0.278)
reg_ecstat	0.082 (0.114)	0.072 (0.114)	0.018 (0.115)
ln_pop	-0.001 (0.03)	0.0003 (0.03)	-0.002 (0.03)
reg_corr	-0.128 (0.296)	-0.07 (0.298)	-0.04 (0.297)
reg_member	**0.886*** (0.329)**	**0.852*** (0.329)**	**0.908*** (0.328)**
reg_christ	**-0.607*** (0.157)**	**-0.597*** (0.157)**	**-0.67*** (0.156)**
reg_icp	**0.983*** (0.277)**	**1.06*** (0.281)**	**0.7** (0.312)**
reg_icp^2		1.874 (1.331)	1.834 (1.351)
reg_ethnic* reg_icp			**-5.769*** (1.637)**
reg_ling* reg_icp			**4.013** (1.649)**
reg_relig* reg_icp			-1.682 (1.188)
reg_politic*			-1.661

reg_icp			(1.332)
k_1	-0.533 (0.467)	**-1.135***** **(-2.86)**	**-1.297***** **(0.336)**
k_2	**1.133**** **(0.467)**	0.53 (0.397)	0.369 (0.336)
k_3	**2.635***** **(0.468)**	**2.032***** **(0.397)**	**1.871***** **(0.336)**
log-likelihood	-27786.876	-27785.89	-27777.562
estimated variance of random intercepts (τ^2)	0.383 (0.042)	0.379 (0.041)	0.349 (0.039)

Note: Standard errors are in parentheses. The number of level 1 units (individuals) is 21642 and the number of level 2 units (regions) is 229 in all model specifications.

Interestingly, none of the coefficients related to the modernization theory have the expected sign. Living in a rural area seems to promote trusting behavior which is a result in line with several previous studies. However, the negative coefficient of education is more unique.[46] The socioeconomic status does not yield a significant coefficient neither at individual nor at regional level. The method as to how the measure is constructed can partly explain this finding. As shown in Appendix 4C, the socioeconomic status is the first principal component that accounts for only slightly more than the half of the variation of the underlying five series. Hence, dimension reduction results in a loss of information that reduces the explanatory power of the composite index. Another possible explanation is related to the content of the underlying queries Q8A to Q8E. This explains how often the respondent has to struggle with the lack of basic necessities, but does not seem to be a close substitute to income and wealth measures to which the modernization theory relates. Hence, according to Model 1, we can at best conclude that the lack of basic necessities is not significantly related to trust. Inequality measured by the standard deviation of the socioeconomic status variable is also insignificant.

The remaining coefficients are predominantly in accordance with our expectations. Dissatisfaction with living conditions and individual corruption perception reduces the log odds of higher trust categories. However, regional level corruption measured with the share of respondents that presume that most police are corrupt is not significant. The importance of voluntary association in trust building is supported at both levels. Population density has no effect.

[46] When we reestimate our models with the original categorical education variable (Q89), we find that people with only informal schooling, completed university and post-graduate education do not show different trusting behavior compared to those with no schooling at all (baseline category). The coefficients of the remaining in-between educational categories (some primary schooling, completed primary schooling, some secondary school/high school, completed secondary or high school, post-secondary qualifications other than university, some university) are negative and significant.

As expected, different aspects of diversity play an important role in trust formation. While religious and political fragmentations seem to reduce the likelihood of trusting behavior, ethnic and linguistic diversity are assigned with positive coefficients. This result supports the contact theory on the relationship between diversity and trust and is in line with the literature that highlights that diversity has different societal impacts in developed and less developed countries. Gundelach (2014) finds that the contact theory is more likely to be relevant in societies where diversity is a stable and historically embedded characteristic of the society and not the results of increasing immigration in the past decades. Also, Anderson and Paskeviciute (2006) argue that ethnic and linguistic diversity erodes civil society in established democracies but not in less democratic countries. However, the coefficient of the linguistic diversity is less robust and becomes insignificant in Model 3.

The results regarding the language related covariates are in line with the first three hypotheses outlined in Section 2. The positive association between languages and generalized trust (Hypothesis 1) is significant only at the level of regions. This finding can be considered as a support for the importance of expectations in the world of imperfect information where the actual individual communication potential is unknown (Hypothesis 2) as well as the role of second languages in counterbalancing the harmful effect of social fragmentations (Hypothesis 3).

The potential non-linearities outlined in Hypothesis 4 are tested in two ways. The aim of introducing the squared term in Model 2 is to detect whether the influence of the regional communication potential is dependent on its value. In Model 3 we consider the interaction of the ICP with the four fragmentation measures. In order to make the estimated coefficients more interpretable and to reduce the possible high correlation between the linear and the squared terms on the one hand and the interaction terms with their component variables on the other, following the common statistical practice, we center the regional ICPs and the fragmentation indices to their mean. Although Model 2 does not confirm the presence of the quadratic effect, certain interaction terms are found to be significant in Model 3. The negative coefficient of the first interaction term suggests that ethnic fragmentation and the communication potential weaken each other's effects. The positive impact of the ICP is smaller in areas with higher ethnic fragmentation and turns negative when the ethnic fragmentation is above 0.721. However, since the variables are centered, the coefficient of the regional ICP indicates that the overall effect is positive in the case of the average person. Similarly, the positive effect of the ethnic fragmentation is weaker in regions with higher communication potential, and the net effect turns negative when the ICP reaches the value of 0.99. The regions that exhibit communication potential over this threshold are located in Botswana, Madagascar, Lesotho and Tanzania, where the extremely high or maximum communication potential is accompanied with low individual variation (Table 4.2). Thus, the net effect of ethnic fragmentation on the generalized trust is positive in countries with an average linguistic situation.

The positive coefficient of the interaction with the linguistic heterogeneity measure suggests that the gains of a unit increase in the regional ICP are larger in regions with higher linguistic diversity. According to the LR test, Model 3 is statistically preferred over both Model 1 and Model 2 at the one percent significance level (the chi-squared statistics are 16.656 with d.f.=4).

The estimated coefficients of the proportional log odds models are translated into probabilities and marginal effects in Table 4.5. All three models suggest that the average individual is most likely to belong to the second trust category (1-just a little). The marginal effects can be interpreted that holding other factors at their mean, increasing the communication potential with an amount would increase the chance of the two upper categories and reduce the likelihood of the two lower answer categories. However, the estimated magnitudes are somewhat smaller in Model 2 and Model 3 compared to Model 1.

Table 4.5 Probabilities and marginal effects at the average individual

		0	1	2	3
Model 1	Probability	24%	39%	26%	12%
	Marginal effect of the reg_icp	-0.17	-0.05	0.13	0.10
Model 2	Probability	23%	38%	27%	12%
	Marginal effect of the reg_icp	-0.19	-0.07	0.08	0.17
Model 3	Probability	21%	38%	28%	14%
	Marginal effect of the reg_icp	-0.1	-0.05	0.08	0.07

Note: Marginal probabilities are in percentage points.

4.4.3 Robustness checks

Lastly, we implement several robustness checks (Table 4.6) to ensure that our results are not sensitive to certain modifications in the model design. First, we replace the ordered categorical dependent variable with a binary choice: the two upper trust categories (I trust them somewhat (2) and I trust them a lot (3)) are recoded into value one; the rest of the answers are coded as zero. This procedure results in a dependent variable that is similar to the traditional trust measure derived from the World Value Survey discussed in Section 4.3.

The second robustness check concerns the possible sensitivity of the results depending on the choice of the sample regions. For the above mentioned reasons we estimate models on a restricted sample excluding countries with extreme communication potential (Botswana, Madagascar, Lesotho, and Tanzania).

The third check focuses on the choice of the trust type as a dependent variable. The dependent variable is now the difference between trust in people the respondents know (Q84B) and the trust in unknown people (Q84C). The coefficients related to ICP from the different alternative specifications are reported in Table 4.6.

Table 4.6 suggests that none of the aforementioned strategies result in qualitative changes in our main results: the beneficial role of the communication potential in trust promotion is confirmed by all specifications. The negative coefficient of the ICP in the last column can be interpreted that the difference between the reported trust in known and unknown people is significantly smaller in geographical regions with higher average communication potential.

Table 4.6 Robustness checks

	Strategy 1 (binary dependent variable)		Strategy 2 (restricted sample)		Strategy 3 (relative trust)
	Model 2	Model 3	Model 2	Model 3	Model 3
reg_icp	**1.377*** (0.312)**	**0.93*** (0.346)**	**1.037*** (0.375)**	**1.004*** (0.377)**	**-0.635** (0.291)**
reg_icp^2	**2.455* (1.464)**	2.254 (1.494)	**3.198* (1.852)**	2.788 (1.84)	-0.818 (1.253)
reg_ethnic* reg_icp		**-5.546*** (1.812)**		-2.98 (2.489)	0.347 (1.53)
reg_ling* reg_icp		**5.193*** (1.814)**		-1.307 (2.381)	-1.418 (1.538)
reg_relig* reg_icp		-1.658 (1.317)		0.357 (2.036)	1.382 (1.1)
reg_politic* reg_icp		**-2.553* (1.478)**		-1.06 (1.943)	2.01 (1.234)
individual variables	yes	yes	yes	yes	yes
regional variables	yes	yes	yes	yes	yes
log-likelihood	-13542.233	-13534.493	-22914.73	-22910.64	-30028.529
estimated variance of random intercepts (τ^2)	0.446 (0.052)	0.41 (0.048)	0.364 (0.047)	0.345 (0.045)	0.292 (0.034)
# of level 1 units	21642	21642	17867	17867	21612
# of level 2 units	229	229	163	163	229

Note: Standard errors are in parentheses. The communication potential and fragmentation measures are centered around their mean.

Although it is not the case in the original model design (Table 4.4), the quadratic relationship is found to be significant when the binary dependent variable is applied. One explanation for this might be that potential measurement errors caused by uncertainties regarding the choice between 0 and 1 on the one hand and 2 and 3 on the other are reduced by collapsing the original four answers into two categories. Measurement errors in the dependent variable, however, do not bias the parameter estimates but reduce the efficiency of the estimator. The quadratic term turns out to be significant at the ten percent level on the restricted sample as well.

The significant quadratic relationship can be viewed as an indication that languages are network goods. The positive coefficient of the squared term suggests that the gains in terms of generalized trust are larger in regions where the average communication potential is higher. The quadratic form also implicates that there is a threshold level under which the increasing communication potential reduces the likelihood of trusting behavior. These threshold levels are 0.453 and 0.52 in the model with the binary dependent variable and the restricted sample, respectively. When the interactions with the four fragmentation measures are taken into account, the squared term becomes insignificant at the traditional levels of statistical significance in both cases. The reduction of the significance of the squared terms can be a sign that the quadratic relationship and interactions capture the same non-linearity in the relationship between languages and trust. Nevertheless, the reduction in the significance is not that remarkable: in both cases the p-value of the coefficient estimates increases from about 0.08 to around 0.13.

Although the interaction terms have not been proven to play an important role in trust shaping in the case of the second and third robustness check strategies, the interaction of the ethnic and linguistic fragmentation with the communication potential still holds in the case of the first strategy; however, the gap between their magnitudes is considerably smaller than in Table 4.4. Moreover, political fragmentation is found to weaken the positive impact of the communication potential in the case of the first strategy.

In the case of the binary dependent variable, Model 3 is preferred over Model 2 at the one percent significance level according to the LR test (chi-squared statistics= 15.48, d.f.=4, p-value=0.004). The better performance of Model 3 compared to Model 2 is confirmed only at the ten percent level when we rely on the restricted sample (chi-squared statistics= 8.18, d.f.=4, p-value=0.085).

This analysis provides robust evidence that the regional communication potential can be considered as an important contributor to social trust. Although there are indications that the effect on trust might be non-linear and might be influenced by certain aspects of diversity, these findings are less stable across model specifications with alternative dependent variables and sample designs.

4.5 Conclusion

In this chapter we test whether languages have a beneficial effect on generalized trust in Sub-Saharan Africa. The linguistic situation is measured with the Index of Communication Potential that can be described as the probability that an individual is able to communicate with another randomly selected person within the society based on common languages. The regional means of this indicator are also used in the empirical analysis. The data are obtained from the fourth round of the Afrobarometer Survey.

Based on theories derived from economics (transaction costs, imperfect information, network externalities, cross-cutting cleavages) and psychology (cognitive biases, heuristics, law of small numbers, outgroup heterogeneity hypothesis, similarity-attraction hypothesis) a simple theoretical framework is established to show the channels through which languages can work and articulate four hypotheses to be tested empirically.

The results of the two-level hierarchical models provide evidence that the ability to communicate with others in a naturally multicultural (multiethnic, multilingual) society promotes social trust (Hypothesis 1). The results indicating that the coefficient of individual ICP is not significant, but that of the regional mean is, is in line with Hypothesis 2. This hypothesis argues that since people are not aware of their actual communication potential, it is rather their expectations about the language repertoire of others that influence their attitudes. Since expectations are based on experience accumulated in everyday life, we assume that regional ICP can be a sufficient measure to proxy individual expectations. The positive coefficient of the regional ICP can be also considered as a support for Hypothesis 3 that second languages can be a sufficient base for cross-cutting cleavages that moderate the harmful effects of fragmentation defined along various dimensions. And finally, although we find possible indications that due to the network effects of languages and the interaction of fragmentation indices with the regional ICPs there might be non-linearities in the relationship between languages and generalized trust (Hypothesis 4), results are not robust across model specifications.

Appendix 4A

Table 4A.1

The number of individuals and geographical regions in the fourth round of the Afrobarometer Survey

country	number of respondents	number of geographical regions	mean number of respondents per region	region with the least observations	region with the most observations
Benin	1200	12	100	64	136
Botswana	1200	24	78	8	144
Burkina Faso	1200	13	92	40	168
Ghana	1200	10	120	40	232
Kenya	1104	8	138	88	264
Lesotho	1200	10	120	48	280
Liberia	1200	15	80	16	392
Madagascar	1350	21	85	10	160
Malawi	1200	3	400	152	560
Mali	1232	9	137	32	216
Mozambique	1200	11	109	64	232
Namibia	1200	13	92	40	208
Nigeria	2324	37	63	14	216
Senegal	1200	11	109	56	256
South Africa	2400	9	267	160	388
Tanzania	1208	26	58	24	104
Uganda	2431	4	608	535	672
Zambia	1200	9	133	72	200
Zimbabwe	1200	10	120	72	200
Total	26449	255	153	-	-

Cape Verde is the twentieth country in the original dataset but we exclude it from our analysis for the reason discussed in the chapter.

Appendix 4B

The construction of the Index of Communication Potential (ICP)

The individual communication potential is defined as the chance that an individual in region r of country k can communicate, i.e. has at least one common language with a randomly chosen individual from the same country. The individual's ability to communicate is recorded in a n_{rk}xn_k matrix M_{rk} (Eq. 4B.1), where n_{rk} denotes the number of individuals in region r (r=1 to R_k) of country k (k=1 to 20), and n_k is the total number of individuals in country k.

$$\mathbf{M_{rk}} = \begin{bmatrix} & 1 & 2 & \cdots & j & \cdots & n_k \\ 1 & m_{11rk} & m_{12rk} & \cdots & m_{1jrk} & \cdots & m_{1n_{rk}rk} \\ 2 & m_{21rk} & m_{22rk} & \cdots & m_{2jrk} & \cdots & m_{2n_krk} \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \vdots \\ i & m_{i1rk} & m_{i2rk} & \cdots & m_{ijrk} & \cdots & m_{in_krk} \\ \vdots & \vdots & \vdots & & \vdots & \ddots & \vdots \\ n_{rk} & m_{n_{rk}1rk} & m_{n_{rk}2rk} & \cdots & m_{n_{rk}jrk} & \cdots & m_{n_{rk}n_krk} \end{bmatrix} \quad \text{(Eq. 4B.1)}$$

If individual i has at least one common language with individual j based on the queries Q3 (What is your home language?) and Q88E (What languages do you speak well?), then m_{ijrk}=1, otherwise 0. The individual index of communication potential can be calculated in both the unweighted (ICP_{irk} , Eq. 4B.2) and weighted (ICP^{w}_{irk} , Eq. 4B.3)) form:

$$ICP_{irk} = \frac{\sum_{j=1}^{n_k} m_{ijrk} - 1}{n_k - 1} \text{ and} \quad \text{(Eq. 4B.2)}$$

$$ICP^{w}_{irk} = \frac{\sum_{j=1}^{n_k} w_{jrk} m_{ijrk} - w_{irk}}{\sum_{j=1}^{n_k} w_{jrk} - w_{irk}} = \frac{\sum_{j=1}^{n_k} w_{jrk} m_{ijrk} - w_{irk}}{n_k - w_{irk}} \quad \text{(Eq. 4B.3)}$$

, where w_{irk} denotes the weight assigned to individual i in region r of country k and $\sum_{i=1}^{n_{rk}} w_{irk} = n_{rk}$, $\sum_{j=1}^{n_k} w_{jrk} = n_k$.

The unweighted and weighted regional ICPs (Eq. 4B.3 and Eq. 4B.4, respectively) are obtained by aggregating the individual ICPs within a region.

$$ICP_{rk} = \frac{\sum_{i=1}^{n_{rk}} ICP_{irk}}{n_{rk}} \text{ and} \qquad \text{(Eq. 4B.4)}$$

$$ICP_{rk}^{w} = \frac{\sum_{i=1}^{n_{rk}} w_{irk} ICP_{irk}^{w}}{\sum_{i=1}^{n_{rk}} w_{irk}} = \frac{\sum_{i=1}^{n_{rk}} w_{irk} ICP_{irk}^{w}}{n_{rk}} \qquad \text{(Eq. 4B.5)}$$

Appendix 4C

The socioeconomic status variable

Since the Afrobarometer Survey does not include questions regarding income or wealth, factor analysis is applied to grasp the socio-economic status of individuals as introduced by Mattes (2008) as the Lived Poverty Index (LPI).

Underlying questions for the socio-economic status variable:

Q8. Over the past year, how often, if ever, have you or anyone in your family gone without:

Q8A. Enough food?
Q8B. Enough clean water from home use?
Q8C. Medicines or medical treatment?
Q8D. Enough fuel to cook your food?
Q8E. A cash income?

Possible answers are (1) Never, (2) Just once or twice, (3) Several times, (4) Many times, (5) Always.

We derive one principal component (eigenvalue is 2.667) that accounts for 53.34% of the variance in the data and computed for 26059 observations.

5 Languages and national identity in Sub-Saharan Africa: a multilevel approach[47]

Abstract

This chapter contributes to the recent strand of the empirical political and economic literature that attempts to reveal the determinants of national identification in Sub-Saharan Africa. Although previous survey-based studies provide evidence that the socio-economic characteristics of individuals, the properties of ethnic groups they belong to, and certain country-level variables influence the probability of having positive attitudes toward the ethnic group or the nation, the role of languages has not been studied in this context yet. Inspired by findings of psycholinguistics and related disciplines, we utilize the fourth round of the Afrobarometer Project (surveyed in 2008 and 2009) to conduct analysis on the possible positive relationship between language knowledge and identification in national versus ethnic terms. We introduce two language-related explanatory variables. First, the Index of Communication Potential (ICP) reflects the probability that an individual can communicate with another randomly selected person within the society relying on commonly spoken languages. Second, we take into account the number of spoken languages in one's repertoire. The multilevel models show that although speaking more than two languages increases the chance of identifying in national compared to ethnic terms, the ICP is not significant in this sense on the whole sample. But, when we consider the nationality of the former colonizers, the ICP exhibits positive relationship with national identification on the sub-sample of the former French colonies.

Keywords: Sub-Saharan Africa, language, communication potential, national identity, multilevel modeling

[47] This chapter is forthcoming as Chapter 7 in Gazzola, M. and Wickström, B-A., eds., The economics of language policy, Cambridge: MIT University Press. The author is grateful to the participants of the Economics of Language Policy session of the CESifo Venice Summer Institute 2013, especially Laura Onofri (Universita Ca Foscari di Venezia, Italy), Michele Gazzola (Humboldt University, Germany) and Bengt-Arne Wickström (Humboldt University, Germany) for their helpful comments on the previous version of this study and Peter Foldvari (Utrecht University, Netherlands) and Pál Czeglédi (University of Debrecen, Hungary).

5.1 Introduction

After gaining independence mostly in the 1950s and 1960s, national integration became a priority in ethnically and linguistically fragmented African countries. Political leaders implemented a whole range of possible tools[48] to strengthen national identity and reduce diversity (partly caused by arbitrary borders set by former colonizers at the Berlin Conference in 1884-85). Due to increased economic and political interest in the past two decades, the societal impacts of ethnolinguistic heterogeneity are well documented.

Although ethnolinguistic fragmentation is traditionally associated with poor economic performance (Alesina et al. 2003, Hall and Jones 1999, Easterly and Levine 1997, Pool 1972), insufficient provision of public goods (Habyarimana et al. 2007, Alesina et al. 1999), higher level of corruption (Mauro 1995), less social trust (Putnam 2007), and higher probability of internal conflict (Montalvo and Reynal-Querol 2005), recent studies highlight that diversity is not harmful for the society by necessity. Treating ethnolinguistic fragmentation as a property of the society that is not unchangeable but determined by historical processes[49], Arcand and Grin (2013) show that it has positive effect on economic development (as measured by GDP per capita) on the sample of a wide range of countries. As a new approach, some scholars suggest that people have to be careful with interpreting cultural and linguistic diversity as a completely harmful condition of certain societies which must be eliminated. It is well established that biodiversity and linguistic diversity are strongly associated (Maffi 2008, Skutnabb-Kangas 2003). Romaine (forthcoming) argues that the preservation of linguistic diversity should be an essential dimension of the 'sustainable development' philosophy since the decline of biodiversity, the endangerment of languages and widespread poverty tend to take place in the same geographical areas.[50]

Although, apparently, it is the two subsequent decades after independence that were engaged in evaluating the success and limitations of nation-building policies (Connor 1972, Rivkin 1969, Deutsch and Foltz 1966, Emerson 1961), the past twenty years have again been showing increasing scientific interest in issues related to ethnicity and nationality in Sub-Saharan Africa. There are at least two main reasons explaining this renewed concern. First, apart from Africa (and Asia) the world can be considered democratic.[51] Limited national integration and the high level of ethnic diversity partly explain this lagging in the democratization process. Second, the recent

---

[48] For a systemic description of the types and efficiency of nation-building policies in post-colonial Africa see Bandyopadhyay and Green (2013).

[49] In statistical terms it means that the authors rely on the instrumental variables (IV) technique instead of the simple OLS method and treat ethnlinguistic diversity as endogenous in their regression analyses.

[50] For a more detailed overview on the negative and positive effects of diversity consult Ginsburgh and Weber (forthcoming).

[51] This statement is based on information provided by the Polity IV Project.
http://www.systemicpeace.org/polity/polity4.htm [15 Aug 2013]

emergence and improvement of regular thematic surveys (World Value Survey, Afrobarometer Survey etc.) provides opportunity to conduct comparative empirical investigations on relevant economic and political issues that have not been feasible earlier due to the lack of appropriate data.

However, it is quite peculiar that after independence when decision upon official languages and language of education have been a crucial part of nation-building policies and the general political discourse in Africa, economic and political research fails to provide comprehensive theory or empirical evidence on the effectiveness of the nation-building tools. There are several possible reasons why the literature has been lacking empirical evaluation on the effects of language policies on ethnic identity at the individual and national integration at the country level. First, identity is a malleable concept that can be defined along different dimensions (e.g. ethnic, social, national identity) that eventually might overlap. Moreover, ethnic identity is not only a part of the broader concept of social identity, but consists of several pillars such as language, religion, cultural norms etc. Definitional vagueness impedes measurement and successful empirical analysis. Second, available language-related data are not applicable for this type of research. In order to analyze the impacts of language on social and political outcomes, we need to separate language from ethnicity. Since ethnicity and language are closely related, they are often used as substitutes or synonyms in the literature. Although the widely known heterogeneity measure introduced by Taylor and Hudson (1972) and used by Easterly and Levine (1997) in their popular work on 'Africa's growth tragedy' is called ethnolinguistic fragmentation (ELF) index, it is based on linguistic data drawn from the Atlas Narodov Mira (1964) compiled by Soviet scholars. Furthermore, it is a common strategy to use linguistic data to proxy ethnic membership when information on the latter is not available (Cheeseman and Ford 2007). Including separate questions on ethnicity and language, as it is done in the Afrobarometer Survey, might solve this type of problem to some extent. However, if we intend to provide a realistic picture on linguistic situation of traditionally multilingual societies, we cannot neglect the importance of second languages. Unfortunately, due to data availability issues, the political and economic literature has been limited to the possibility of analyzing home languages and ethnolinguistic fragmentation in Sub-Saharan Africa. Investigation on second languages might open new research directions in development and political studies. Language learning might partially balance out the detrimental effects of ethnolinguistic fragmentation by reducing transaction costs of communication and cooperation. In addition, language learning can be seen as investment in 'similarity' which is a crucial determinant of interpersonal trust and willingness to cooperate with others (Leeson 2005).

The main contribution of this chapter is that, as to our knowledge, this is the first attempt to provide empirical evidence on the possible relationship between language knowledge and ethnic or national identification on a broader sample of about 27,000 individuals in twenty Sub-Saharan African countries. Psycholinguistic research and a

few economic papers related to this issue (see the next section) are usually based on case studies conducted among immigrants in developed countries (Netherlands and the United States) or bilingual people in certain societies (Canada and Spain). This study tests if the relationship between language and identity experienced in bilingual or immigrant communities still exists in the multilingual setting of Africa.

To reveal the relationship between language knowledge and the importance of nationality compared to ethnicity, we use the database of the fourth round of the Afrobarometer Survey. We create two measures that are supposed to grasp the individual-level language knowledge. The Index of Communication Potential (ICP) is the probability that a person might communicate with another randomly selected individual within the society based on common languages. In addition, we control for the number of spoken languages. We estimate multilevel logit models that are designed to analyze hierarchical data, i. e. observations at the basic level of analysis (in our case individuals) might be organized into higher level units (in our case ethnic groups).[52] Thus, we are able to handle statistically that individuals within an ethnic group are more similar than across ethnic groups. At the level of individuals, we include common socio-demographic variables such as age, gender, location of residence, education, socio-economic status, religion, and language abilities in the model. At the ethnic group level, we assume that the intercepts of the regression lines might vary randomly among ethnic groups. While speaking more than two languages contributes to higher national identification on the whole sample, we find no such a significant effect when individual language abilities are measured with the ICP. Although, when we experiment on various sub-samples defined by colonial history, we find empirical evidence that the Index of Communication Potential is associated with stronger national feelings on the sub-sample of former French colonies. However, languages do not seem to be relevant in identity shaping on the British territories at all.[53]

The chapter proceeds as follows. The next section provides literature background referring to psychological, sociolinguistic, political and empirical economic studies that inspired this work. Section 3 describes our data and variables used in the empirical analysis. Section 4 is devoted to the discussion of the multilevel analysis on the total sample and sub-samples. The final section concludes and articulates further research steps.

[52] The benefits of this method are discussed in more detail later in the chapter.

[53] The author is aware of the possibility of reverse causality between language knowledge and identity (not only languages might influence feelings toward the nation or the ethnic group but the relationship is possible to exist in the other direction as well). But since the aim of this chapter is to show the expected positive relationship between languages and national feelings, we do not address this issue in detail. Moreover, our method that is able to handle the hierachical structure of our dataset is not easily applicable to treat the statistical consequences of reverse causality.

5.2 Related literature

The aim of this section is to give an overview on the literature that has inspired this present study. Theoretical discussions and case studies provided by social sciences suggest that language, ethnicity and identity are interrelated. We attempt to integrate this broad idea into the line of the largely empirical political and economic literature that seeks to reveal the determinants of ethnic and national identification in Sub-Saharan Africa.

Revealing the relationship between language, ethnicity (race) and identity has long been the concern of psycholinguistics (neurolinguistics and second language acquisition studies (SLA)), sociolinguistics, linguistic anthropology, and political studies.[54] These fields have attempted to define what should one mean by 'identity' and 'ethnicity' and what is included in those terms (Chandra 2006, Hale 2004, Sellers et al. 1998). One of the most debated issues is the origin of ethnic groups: are they historical entities shaped by 'natural' long-lasting processes or created by the activity of third parties (such as colonizers or missionaries) or modern phenomena (Bayar 2009, Brubaker et al. 2004, Fishman 2002) [55]. Some studies claim that ethnic identity is not a stable or an absolute construction but might be dynamic and determined by situational factors (Yip 2005, Yip and Fuligni 2002, Shelton and Sellers 2000, Kaufert 1977, Nagata 1974). Other considerations are related to the link between ethnicity and national identity (Chee-Beng 2000, Hutchinson 2000, Smith 1986, Gellner 1983) and the role of languages in the construction of identity (Bucholtz and Hall 2010, Johnstone 2010, Simpson 2008, Joseph 2004). Finally, recent language planning and policy questions are raised not only in relation to the natural multilingual settings in the less developed world (Africa or Asia) but the increased diversity in developed countries due to immigration (King and Rambow 2012). Moreover, the enlarged European Union has witnessed the emergence of some crucial supranational linguistic issues in the past decades (Ginsburgh and Weber (forthcoming), Ammon 2012, Fidrmuc et al. 2009, Phillipson 2006).

Turning to the field of political science, Laitin provides two monographs on the relationship between language, political culture and identity. In an early book (Laitin 1977), he notes that although socio- and psycholinguistics suggest that language might help form and maintain cultural norms, evidence that supports these assumptions is missing. He elaborates on the case of Somalia (a country that had been exposed to relevant foreign linguistic influence) to show how language issues might be related to politically relevant phenomena such as political participation and political thoughts. In

[54] For a general summary on the approaches of different social sciences to language and ethnic identity see Fishman (1999).
[55] This is the so called 'primordialist vs. constructivist' debate.

a later work (Laitin 1998), he concerns the Russian-speaking population in the post-Soviet republics that face identity crisis after the collapse of the Soviet Union.

Analyzing issues related to identity is not the exclusive privilege of psychology, anthropology and sociology, economics also has the potential. In a groundbreaking paper, using a basic game theoretic model, Akerlof and Kranton (2000) show how theories explaining the presence of different socio-economic phenomena (gender discrimination at workplace, poverty and social exclusion, and the division of labor at the household level) might benefit from incorporating identity (defined as the person's sense of self) in the utility function. In their view, identity is a complex structure that is associated with different social categories. Bodenhorn and Ruebeck (2003) focus solely on the ethnic and racial component of identity. Their main assumption is that ethnic and racial identity is a fluid concept and might be dependent on individual choices influenced by economic factors. Using the case of the free African-American population in the mid-nineteenth century, they investigate the costs and benefits of maintaining or changing racial identity. Since the above discussed seminal paper of Akerlof and Kranton (2000), several studies have contributed to the field of 'identity economics' by addressing questions regarding the concept of identity (Fine 2009 and Hill 2007) and discovering additional application possibilities (Akerlof and Kranton 2010, 2005, and 2002, Davis 2010, and 2007, Darity Jr. et al. 2006). However, these papers do not propose a possibility to incorporate language into the framework.

Regarding the empirical literature, we find only a few economic studies on testing the relationship between linguistic issues (language policy and second language learning) and identity formation. In a paper series on introducing bilingual education in Catalonia and the Basque Country in Spain in 1983, we find empirical evidence that this type of relationship might exist and can be measured. Clots-Figueras and Masella (2013) show that individuals who have experienced more exposure to Catalan language at school tend to feel more Catalan than Spanish. This result persists among pupils whose parents do not have Catalan origins. Moreover, this identity change influenced by policy and language learning might impact political behavior as well. People in Catalonia who have experienced more teaching in Catalan are more likely to declare to have voted for a Catalanist party in 1999. Aspachs-Bracons et al. (2008) also find evidence that it is not only the introduction of the bilingual education what mattered in shaping identity but the fact whether having education in Catalan and Basque was compulsory or optional. In Catalonia, where the introduction of the local language in education was compulsory, they find that more schooling in Catalan increased the chance to feel more Catalan than Spanish. In the Basque country, where education in Basque was optional, they do not document such an effect.

A number of empirical works exploit the rich database of the Afrobarometer Survey (more on the database in Section 5.3) and examine the potential determinants of ethnic or national identification in Sub-Saharan Africa. These factors are derived from various theoretical considerations and can be measured at the country, ethnic group and individual level. Most scholars agree that nationalism is a modern phenomenon

that had emerged in the nineteenth century as a result of wars, industrialization, print capitalism, and strategic state policies. This implies that characteristics of a modern society (more education, formal sector employment, urbanization, industrialization) are expected to be associated with higher level of national identification (Robinson 2009). Although Robinson (2009) provides empirical support for this theory, Bannon et al. (2004) reveal opposite relationship and find that more education, non-traditional occupation, and living in an urban area are likely to increase the probability of identifying in ethnic terms.

Colonial history might also have impact on the contemporary level of nationalism experienced in a society. Since, unlike the other European imperia (French, Belgian, and Portuguese), they acknowledged local tradition and social organization to a greater extent and were not engaged in assimilating local people into the broader metropolitan culture (Lange 2004, White 1996), former British colonies are expected to exhibit less nationalism.

The ethnic composition of a society is an additional potential factor that influences national integration. Ethnic diversity is assumed to impede successful nation-building efforts. There is empirical evidence that African countries with higher level of ethnic fragmentation show less national sentiment (Bannon et al. 2004). However, it is also possible that the relative size of an ethnic group within a society influence the behavior of group members. Posner (2004b) shows that the Chewa and Tumbuka communities are more engaged in political affairs in Malawi than in Zambia. While in Malawi, these two ethnic groups are relatively large and provide a potential basis for political coalition-building, in Zambia, due to their relative small size, Chewas and Tumbukas are not useful to mobilize as bases of political support. Robinson (2009) points out that the relative size of ethnic groups within the society and the extent to which ethnic groups were partitioned by the colonial borders have significant impact on the national sentiment. Larger ethnic groups, groups that are not divided between countries, and groups that have large proportion outside the country tend to feel more national.

Several studies show that identity is not a stable construction but might be dependent on situational factors. Eifert et al. (2010) argue that African people tend to think in ethnic terms when political elections are approaching. This means that ethnicity has instrumental relevance in Africa: it is a useful concept in competition for political power.

Besides the above mentioned potential factors, most empirical research control for individual demographical characteristics of individuals (age, gender, occupation, years of education, location of residence, religion), but their effect on identity is varying among studies.

5.3 Data and variables[56]

Similarly to previous papers in the field, this study relies on the Afrobarometer Survey[57]. The Afrobarometer is considered a politically independent project aiming to map the political atmosphere in Africa. The advantage of this source is that surveys are repeated on a regular cycle[58] and ask a standard set of questions which makes comparative research possible not only across countries but even over time. Due to this benefit, the dataset is becoming more and more popular in economic and political research (Michalopoulos and Papaioannou 2013, Nunn 2010, Kaufman et al. 2004). The sample in the Afrobarometer Project is designed as a representative cross-section of all citizens of voting age in a given country.[59] Since information on second languages is included only in Round 4 (conducted in 2008 and 2009), we have to restrict our analysis to that round. The original source covers twenty Sub-Saharan African countries, but for several reasons (detailed later) we exclude six of them from some model specifications (Table 5.1).

Table 5.1 Sub-Saharan African countries in the 4th round of the Afrobarometer dataset

country	sample size	in all specifications	country	sample size	in all specifications
Benin	1200	yes	Mali	1232	yes
Botswana	1200	no	Mozambique	1200	yes
Burkina Faso	1200	yes	Namibia	1200	yes
Cape Verde	1264	not included	Nigeria	2324	yes
Ghana	1200	yes	Senegal	1200	yes
Kenya	1104	yes	South Africa	2400	yes
Lesotho	1200	no	Tanzania	1208	no
Liberia	1200	yes	Uganda	2431	not included
Madagascar	1350	no	Zambia	1200	yes
Malawi	1200	yes	Zimbabwe	1200	yes

5.3.1 Dependent variable – the relative importance of national compared to ethnic identity

The dependent variable used in the empirical investigation is supposed to refer to the strength of national identity. Given the available survey and following the 'tradition'

[56] Variables described in this section and used in the empirical analysis are detailed in Appendix 5A.
[57] www.afrobarometer.org
[58] Round 1 conducted between 1999 and 2001 with 12 countries included, Round 2 between 2002 and 2004 with 16 countries included, Round 3 in 2005 and 2006 with 18 countries included and Round 4 in 2008 and 2009 with 20 countries included. Round 5 that covers 36 countries now including Northern African countries is being processed and digitalized at the moment.
[59] 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 into the sample.

established by the related empirical literature, we utilize the following question of the 4th round of the Afrobarometer Survey that measures the relative importance of nationality compared to ethnicity:

Q83. Let us suppose that you had to choose between being a [Ghanaian/Kenyan/etc.] and being a [Respondent's ethnic group]. Which of the following best expresses your feelings?[60]

The possible answers:

(1) I feel only (Respondent (R)'s ethnic group)
(2) I feel more (R's ethnic group) than [Ghanaian/Kenyan/etc.]
(3) I feel equally [Ghanaian/Kenyan/etc.] and (R's ethnic group)
(4) I feel more [Ghanaian/Kenyan/etc.] than (R's ethnic group)
(5) I feel only [Ghanaian/Kenyan/etc.]
(7) Not applicable
(9) Don't know
(998) Refused to answer

Possible answers (7), (9), (998), and missing data are excluded from the final analysis. The remaining answers are measured on a five-point ordinal scale where the value of one suggests that the respondent does not feel to belong to the nation at all, while the value of five refers to the highest level of national commitment compared to ethnic membership. Figure 5.1 shows the distribution of answers to Q83 in our twenty sample countries. The general pattern is that answer (3) and answer (5) are the most dominant. We find that in Ghana, Lesotho, Liberia, Namibia, Nigeria, Uganda, Zambia, and Zimbabwe the 'neutral' answer (3) is the most frequent answer category, while in Cape Verde, Madagascar, Malawi, Mali, Senegal, South Africa, and Tanzania people rather feel national. Benin, Botswana, Burkina Faso, Kenya, and Mozambique exhibit a special pattern with answer (3) and (5) chosen almost equally frequent.

[60] This identity-related question might be criticized from different point of views (Robinson 2009). First, it is unable to take into account situational factors that might influence the answering process to a question. Secondly, this certain type of question investigates the importance of national compared to ethnic identity, thus implicitly assumes that national and ethnic identity are two edges of the same scale, i.e. exclude each other. The first two rounds of the Afrobarometer Survey encountered an identity related question of a different nature: *We have spoken to many [people in this country] and they have all described themselves in different ways. Some people describe themselves in terms of their language, religion, race, and others describe themselves in economic terms, such as working class, middle class, or a farmer. Besides being a [citizen of this country], which specific group do you feel you belong the first and foremost? Possible answers: (0) Can't explain, (1) Language/tribe/ethnic group, (2) Race, (3) Region, (4) Religion, (5) Occupation, (6) Class, (7) Gender, (8) Individual/personal, (10) Won't differentiate/national identity, (12) Traditional leader, (13) Political party identity, (14) Age-related, (15) African/West African/Pan African, (16) Island, (50) Portuguese, (51) American, (60) Family/relationship-based (e.g. wife, parent, widow, etc.) (61) Marginalized group (e.g. disables, etc.), (995) Other, (999) Don't know.* A handbook on the difficulties of measuring identity is provided by Abdelal et al. (2009).

A variable like Q83 provides us with several strategies to implement empirical analysis. In the current form, this variable might be applied in an ordered logistic model. The drawback of this strategy is that it estimates only one coefficient for each explanatory variable which means that a significant covariate increases or decreases the chance of a higher-level answer category always the same way no matter what the initial category is. Another possibility is to estimate a multinomial model that handles each outcome as discrete instead of ordered, and the probability of each outcome is compared against the selected reference category. And finally, a binary variable might also be created. However, transforming a variable measured on an ordinal scale into a binary one might raise two types of problems. First, recoding might be arbitrary, which can influence the regression results. Second, some information on the original structure and the distribution of answers could be lost in the transformation process.

Due to the special nature of answer (3) as a neutral answer, we create two binary variables as potential dependent variable in the following empirical analysis. In both cases answer (1) and (2) are recoded into zero, and answer (4) and (5) are recoded into one. In the recoded variable 'id_bin' answer (3) is assigned with the value of zero, while in the case of 'id_bin2' answer (3) is assigned with the value of one. The difference between the two versions is shown in Figure 5.2 and Figure 5.3. Since the 'id_bin' binary variable in Panel A exhibits more variation, thus contains more information, we utilize it in all specifications of the logistic multilevel models in the next section.

Finally, as an interesting point, we have to remark that excluded answers (7) and (9) would provide an opportunity to study the importance of ethnicity from a special angle. 'Not applicable' in Q83 means that when the respondents are asked about their ethnic group in Q79[61], the answer is 'Ghanaian/Kenyan/etc. only', 'does not think in those terms', 'Refused to answer' or 'Don't know'. Not thinking in ethnic terms might also be exploitable for our purposes. Since it suggests that the respondent does not have strong revealed attitudes neither toward the ethnic group nor the nation, the answer 'Don't know' to Q83 could have been treated in a different way than simply ignoring it. This strategy, however, would raise two questions: where to put answer (9) in the order if we apply the ordered version of the variable, and how to code it when we intend to use the binary version of the dependent variable.

The distribution of respondents that do not know their ethnic group or does not think in ethnic terms is presented in Figure 5.4. The most striking case is Cape Verde with the most respondents that do not know their ethnic groups and do not think in those terms. In Mozambique, 15% of the respondents do not know their ethnic group, and in South Africa about 12% does not think in ethnic terms.

[61] The question regarding the ethnic group is the following in each country. *Q79. What is your tribe? You know, your ethnic or cultural group.* The possible answers list the most representative ethnic groups in each country. Further possible answers are (990) 'Ghanaian/Kenyan/etc. only or does not think in those terms', (995) Others, (998) Refused to answer, (999) Don't know.

Figure 5.1 The distribution of answers to Q83 in the twenty SSA countries

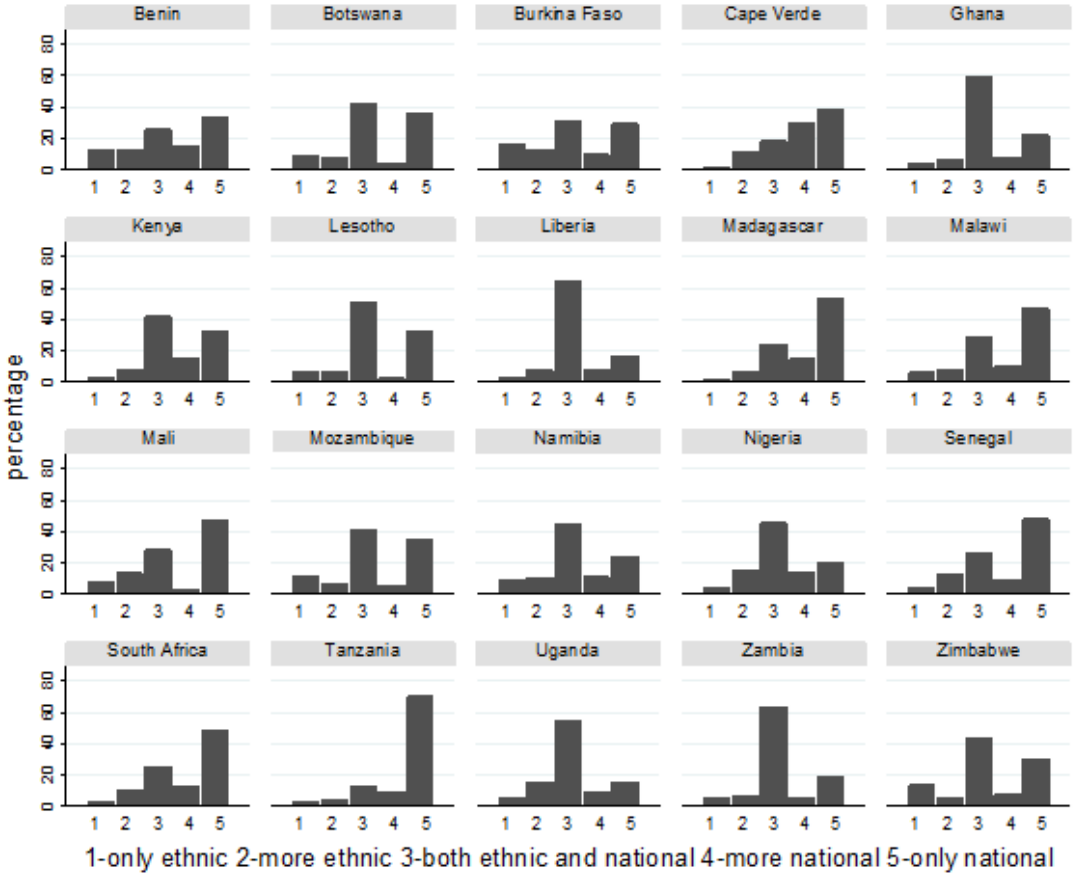


Figure 5.2 The share of respondents with national identity more important as ethnic identity

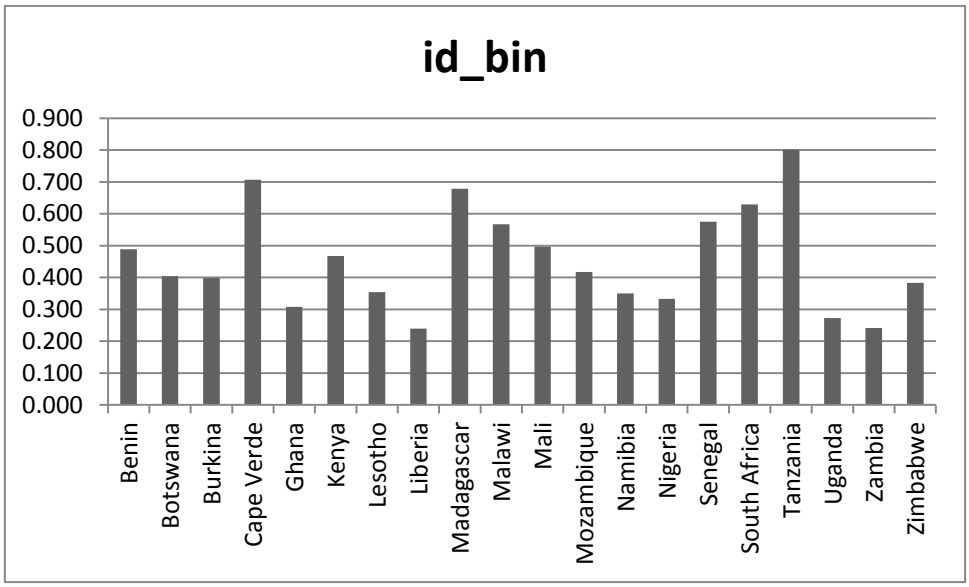


Figure 5.3 The share of respondents with national identity at least as important as the ethnic identity

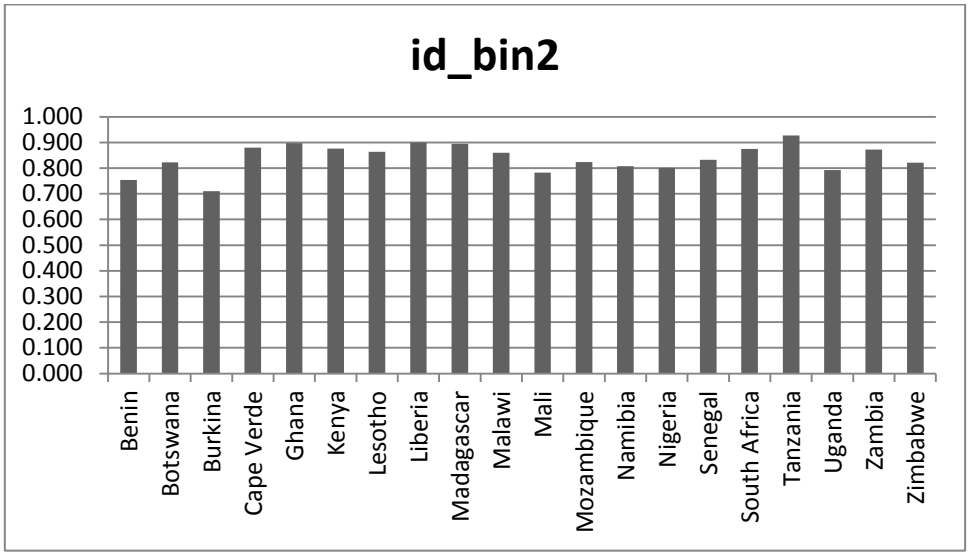
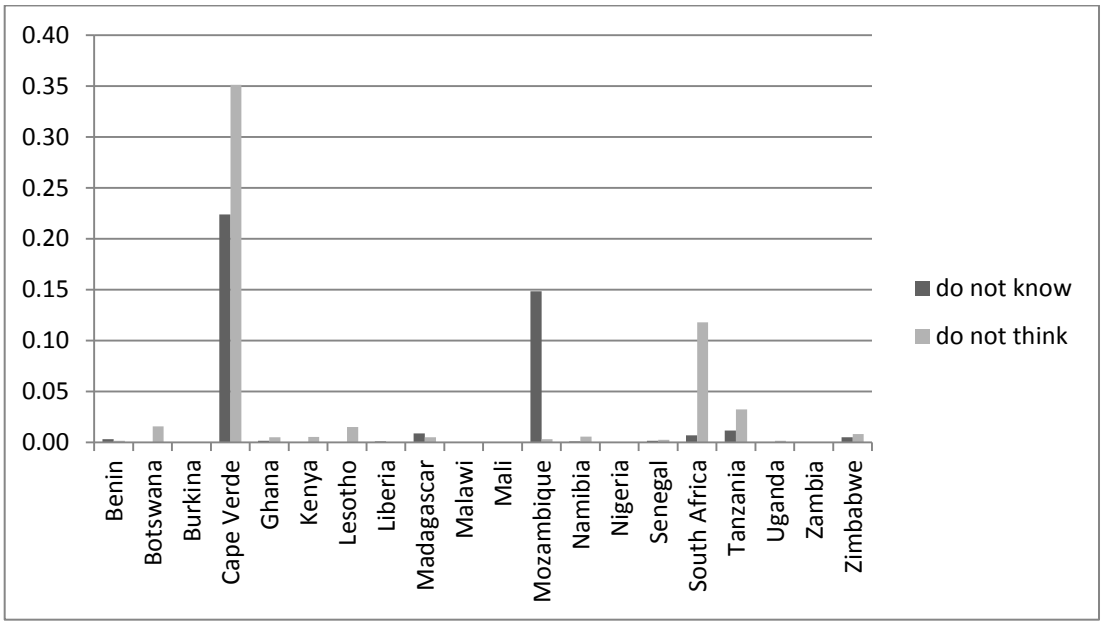


Figure 5.4 The distribution of respondents that do not know their ethnic group or do not think in ethnic terms



5.3.2 Main explanatory variables - the Index of Communication Potential (ICP) and the number of spoken languages

In order to shed light on the possible relationship between language knowledge and national identity, we have to construct appropriate language-related variables. Based on Q3 (Which Ghanaian/Kenyan/etc. language is your home language?) and Q88E (What languages do you speak well?), we experiment with two measures to capture the language knowledge of individuals covered by the survey.

The Index of Communication Potential (ICP)[62] refers to the probability that a person can communicate with a randomly selected other individual in the country based on common languages. Technical details are presented in Appendix 5B. It is important to notice here that the ICP captures the communication possibilities in the simplest form possible, i.e. beyond common languages it does not include other potential factors that might influence communication within the society. The ICP is not designed to account for the disinclination between ethnic and linguistic groups either. In this current form, it is neither able to consider the possible consequences of the geographical distribution of ethnic and linguistic groups. If members of different ethnic and linguistic groups live next to each other, they are more encouraged to find the way to communicate. Nation-building is likely to be less difficult in a country where groups are not separated geographically. Although the Index of Communication Potential might be redesigned to take all the aforementioned dimensions into account, we leave it to a later research.

The second measure that is designed to grasp someone's language knowledge is the simple number of spoken languages. However, we assume that the possible effect of an additional language is dependent on the initial number of mastered languages, i.e. the effect of the second language on identity is expected to be higher than the effect of the third, fourth, etc. language. Therefore, we categorize respondents into groups based on this assumption. The first group consists of those who speak only their home language. Individuals with two languages are the second group. Group 3 includes people who master at least three languages. A country-level summary on the language-related variables is presented in Table 5.2.

Other than home languages are not surveyed in Cape Verde. However, since almost all respondents chose Cape Verdean Creole as home language (and a few named Portuguese), the Communication Potential is very close to one. In Botswana, Lesotho, Madagascar, and Tanzania the situation is similar: high Index of Communication Potential with low level of individual standard deviation (less than ten percent). This is the reason why these above mentioned five countries are excluded from some specifications in the empirical analysis. Uganda is ignored for more practical reasons. The online dataset that is available for Round 4 has changed since the time of the computation of the ICP measures (the number of respondents has changed from 2431 to 4096).

---

[62] This index is different from the Q-value of communication potential introduced by De Swaan (2001 and 1993). Although both indicators attempt to measure the value of language repertoires in terms of the share of a population that can be reached through them, their main aim and construction are different. 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. The Q-value is the product of the 'prevalence' (the proportion of native or foreign speakers of a certain language among all speakers) of a given language and its 'centrality' (the proportion of multilingual speakers of a certain language among all multilingual speakers). As it is shown in Appendix 5B, the ICP is calculated as the simple (weighted) share of the society that an individual can communicate with based on his or her language repertoire.

Table 5.2 Language-related variables in the sample countries

	average number of spoken languages	share of individuals speaking only their home languages (%)	share of individuals speaking two languages (%)	ICP	largest language (first and second speakers)
Benin	2.32 (1.14)	24.54	38.33	0.581 (0.219)	Fon: 59.2%
Botswana	1.99 (1.03)	38.42	38.00	0.986 (0.078)	Tswana: 99.2%
Burkina Faso	1.95 (1.01)	40.10	34.92	0.602 (0.239)	Moore/Mossi: 67%
Cape Verde	-	-	-	0.994 (0.056)	-
Ghana	2.60 (1.18)	18.11	30.67	0.614 (0.225)	Akan: 55.6%
Kenya	2.72 (0.785)	4.51	28.58	0.917 (0.172)	Swahili: 94.3%
Lesotho	1.59 (1.03)	62.91	26.17	1.000 (0.000)	Sesotho: 100%
Liberia	2.04 (0.810)	25.19	51.00	0.598 (0.278)	English: 72.4%
Madagascar	1.30 (0.59)	75.33	20.00	1.000 (0.000)	Malagasy: 100%
Malawi	2.00 (1.000)	35.78	41.25	0.884 (0.200)	Chewa: 92.9%
Mali	2.03 (1.07)	36.72	38.80	0.803 (0.238)	Bambara: 87.2%
Mozambique	2.32 (1.01)	18.43	49.75	0.697 (0.260)	Portuguese: 78.7%
Namibia	2.72 (1.13)	10.69	37.17	0.816 (0.166)	English: 76.2%
Nigeria	2.20 (0.94)	21.98	45.31	0.622 (0.205)	English: 64.5%
Senegal	1.96 (1.03)	38.57	36.17	0.892 (0.179)	Wolof: 92.3%
South Africa	2.26 (1.22)	27.04	45.38	0.606 (0.241)	English: 65.6%
Tanzania	2.26 (0.80)	10.84	60.43	0.991 (0.063)	Swahili: 99.4%
Zambia	2.80 (1.45)	19.92	27.75	0.643 (0.230)	Bembe: 63.9%
Zimbabwe	1.98 (0.97)	35.38	41.25	0.871 (0.183)	Shona: 90%
Mean	2.17 (1.01)	30.25	38.39	0.796 (0.17)	-

Note: Uganda is excluded for reasons detailed in text. Individual standard deviations are reported in parentheses.

The first column of Table 5.2 shows that a typical person in the sample countries speaks more than one language with an average of about two. The standard deviations presented in parentheses show the average individual deviation from the country mean. Low standard deviation indicates that the individual numbers of spoken

languages tend to be close to the country mean, while high standard deviation suggests that the individual numbers of languages are spread out over a larger range of values.

The share of individuals that are proficient in their home language only varies between 4.51 % in Kenya to 75.33% in Madagascar. The corresponding number regarding people with only two languages is between 26.17% in Lesotho and 60.43% in Tanzania. More than 50% of the respondents master at least three languages in Ghana, Kenya, Namibia and Zambia. In countries where there is a recognized national language spoken by everyone or almost everyone (Cape Verde, Lesotho, Madagascar, Tanzania), the ICP is 1 or very close to that. An additional important characteristic of these countries is that at least 90% of the respondents in the sample speak maximum two languages. The last column presents the largest languages either first or second. Tswana contributes to very high ICP in Botswana. The same role is fulfilled by Swahili in Kenya and Tanzania, Chewa in Malawi, Wolof in Senegal and Shona in Zimbabwe.

Finally, certain countries are suggested for further linguistic analysis. As mentioned earlier and shown in Figure 5.1 and 5.2, Tanzania is considered as the most successful case of nation-building policies in Africa. One pillar of the policy package was the introduction of Swahili as the official language, which seems to be the reason for the extremely high Index of Communication Potential. Cape Verde and Madagascar, where the ICP is close to one, are the second and third most 'nationalist' countries within the sample (see Figure 5.2 and 5.3). Botswana and Lesotho produce a somewhat different pattern with the neutral answer (3) as the more frequent.

5.3.3 Other individual-level explanatory variables

Our models control for some demographic characteristics of individuals, such as age, gender, location of residence (urban or rural), years of education, religion, and economic status. Summary statistics on the individual level variables are shown in Table 5.3.

Since the education systems and the duration of school levels show high degree of diversity in Africa, we convert categorical variables (Q89) into average years of education[63]. However, rescaling has induced some confusing questions. How many average years of education should be given to 'started but not completed' categories? Most African countries maintain a divided system of secondary schooling with lower (junior) and upper (senior) levels. Does 'completed secondary education' mean finished junior and senior school both or the lower level only? Further, 'post-graduated' category might mean master and PhD studies. And finally, handling the 'post-secondary' category is also challenging. Post-secondary education is a broad category with an unclear vocational training (e.g. nurse or teacher education)

[63] To find out the duration of the primary and secondary education we applied the Foreign Credit Class Base Education Database (http://www.classbase.com/) [25 Feb 2012]. To get duration of higher level education programs we browsed the websites of main universities of each country.

embraced by this expression. The problem is that some types of vocational education are tied to finished lower and higher level of secondary schooling, while some are based on the lower level only.

Table 5.3 Individual level variables

	average age (in years)	share of men (%)	rural (%)	average education (in years)	muslim (%)	christian (%)
Benin	35.42 (12.95)	50.22	58.57	4.67 (4.84)	25.40	54.82
Botswana	40.03 (17.37)	49.99	76.53	7.59 (4.83)	0.51	50.02
Burkina Faso	36.61 (14.98)	49.94	73.48	3.11 (4.21)	56.84	35.56
Ghana	38.69 (16.12)	50.00	56.08	6.18 (4.87)	15.44	78.68
Kenya	35.36 (12.74)	50.21	79.43	8.68 (3.75)	9.21	86.34
Lesotho	41.14 (18.28)	49.70	69.99	6.33 (3.95)	0	97.41
Liberia	36.27 (12.99)	50.06	52.71	6.48 (4.95)	9.57	87.25
Madagascar	39.66 (14.18)	50.12	74.48	5.93 (4.05)	3.02	62.61
Malawi	35.49 (13.84)	49.88	85.42	5.54 (3.91)	10.93	78.45
Mali	39.15 (14.00)	50.05	73.16	2.18 (3.52)	88.04	4.09
Mozambique	30.81 (12.14)	49.92	68.07	5.46 (3.31)	22.50	60.45
Namibia	34.86 (14.29)	50.26	63.87	9.20 (3.72)	0	96.83
Nigeria	31.40 (11.57)	49.73	93.38	9.26 (4.49)	41.29	56.07
Senegal	39.03 (14.81)	50.73	54.62	4.03 (4.25)	95.77	3.45
South Africa	37.93 (15.31)	53.35	37.70	9.52 (3.59)	2.00	64.07
Tanzania	37.51 (14.05)	49.77	76.31	7.18 (3.13)	28.89	63.18
Zambia	34.91 (13.82)	50.00	63.34	7.52 (3.86)	0.06	84.84
Zimbabwe	36.19 (15.12)	49.99	64.72	9.15 (4.12)	0.34	85.02
Mean	36.69 (14.36)	50.22	67.88	6.56 (4.08)	22.77	63.84

Note: Cape Verde and Uganda are excluded from the empirical analysis. For explanation see text. In the case of age and education the average values are reported, the individual level standard deviations are shown in parentheses. The table excludes the principal component of the socio-economic status which is discussed in Appendix 5C.

In order to overcome the above mentioned difficulties, we apply two rescaling concepts differing in the recognized years of secondary education. The first version assumes that 'finished secondary education' refers to the lower level only, the second

version assumes that both levels are accomplished. In the final empirical analysis, we use the average of the two alternative measures of the years of education. The 'in between' or 'not finished categories' get the average years of the finished categories above and below. Average years of post- secondary schooling equals with the average years of the maximalist version of secondary education. The average year of post-graduation is the average years of the master programs plus one year to control for possible PhD education.

Since the Afrobarometer Survey does not cover questions related to income or wealth, we apply principal component analysis on appropriate survey questions (Q8A, Q8B, Q8C, Q8D, and Q8E) to reveal the socio-economic status of individuals. The principal component analysis requires standardized data (all the factors have zero mean and unit variance). Technical details are reported in Appendix 5C.

5.4 Empirical analysis

This chapter exploits the methodological benefits of the multilevel analysis technique[64] to shed light on the possible relationship between language knowledge (measured with the Index of Communication Potential and the number of spoken languages) and national relative to ethnic identity.

Multilevel modelling technique is suggested when the dataset is hierarchical, i.e. the observations can be classified into higher level units (in our case individuals into ethnic groups). Since similarity within ethnic groups is higher than across ethnic groups, our observations cannot be considered independent. Consequently, the basic requirement for a proper traditional ordinary least squares (OLS) method is violated. The multilevel models might be conceptualized as a multistage system of equations in which individual level variation in each group is explained by an individual-level equation, and the variation across groups is captured by a group-level equation (Diez-Roux 2000).

Although it would be insightful to estimate a three-level model with explanatory variables measured at the level of individuals, ethnic groups and countries, we apply only the two-level model with random intercepts at the level of ethnic groups without ethnic group-level explanatory variables. The limited number of countries (18 and 14) impedes the inclusion of a third level of countries in the model. Our aim is only to provide indication that the relationship between language knowledge and identity exists in the multilingual context of Sub-Saharan Africa. Extending the spectrum of explanatory variables and building more complex models is the subject of a forthcoming study. We estimate the following model (Eq. 5.1, Eq. 5.2, Eq. 5.3):

[64] Technical details on the method of the multilevel analysis are found for instance in Snijders and Bosker (1999).

Level 1 (individual) equation:

$$identity_{ij} = \log(\frac{P(national_{ij}=1)}{P(ethnic_{ij}=1)}) = \beta_{0j} + \beta_{1j} age_{ij} + \beta_{2j} female_{ij} + \beta_{3j} rural_{ij} + \beta_{4j} chris_{ij} + \beta_{5j} educ_{ij} + \beta_{6j} ecstat_{ij} + \beta_{7j} language_{ij} + \varepsilon_{ij} \quad \text{(Eq. 5.1)}$$

Level 2 (ethnic group) equation:

$$\beta_{0j} = \gamma_{00} + \nu_{0j} \quad \text{(Eq 5.2)}$$

The estimated model:

$$identity_{ij} = \log(\frac{P(national_{ij}=1)}{P(ethnic_{ij}=1)}) = \gamma_{00} + \beta_{1j} age_{ij} + \beta_{2j} female_{ij} + \beta_{3j} rural_{ij} + \beta_{4j} chris_{ij} + \beta_{5j} educ_{ij} + \beta_{6j} ecstat_{ij} + \beta_{7j} language_{ij} + \nu_{0j} + \varepsilon_{ij} \quad \text{(Eq 5.3)}$$

where $identity_{ij}$ is the log odds ratio that individual *i* in ethnic group *j* feels more national than ethnic (answer (3) in Q83 is coded as 0 as presented in Section 5.3.1 and Figure 5.2). Age_{ij} refers to the age of individual *i* of ethnic group *j* in years. $Female_{ij}$, $rural_{ij}$ and $chris_{ij}$ are additional demographic dummy variables assuming the value of one if respondent *i* of ethnic group *j* is female, lives in a rural area and Christian, respectively. The $educ_{ij}$ refers to the years spent in education as discussed above and summarized in Table 5.3. The $ecstat_{ij}$ proxies the socio-economic status of individual *i* of ethnic group *j* as discussed in Appendix 5C. The $language_{ij}$ is one of the possible individual level language-related variables (ICP or the dummies derived from the number of spoken languages). ε_{ij} is the individual error term with $\varepsilon_{ij} \sim N(0, \sigma^2)$ distribution, and ν_{0j} is the random part of the intercepts at ethnic group level with $\nu_{0j} \sim N(0, \tau^2)$ distribution. Thus, we assume that the intercept might vary randomly between ethnic groups.

However, we have to restrict our sample to some extent. There are some ethnic groups that have only a few observations. Thus, we include only those ethnic groups in the final analysis that represents at least 3 percent of the sample in each country or have at least 30 observations. Appendix 5D provides detailed information on the incorporated ethnic groups.

Table 5.4 shows the results of the multilevel analysis on the whole sample. Model 1 and Model 2 exploit the ICP as the main language-related variable, while Model 3 and Model 4 make use of the number of spoken languages. Since Model 1 and Model 3 contains 18 countries including those with nearly maximum communication potential and very low individual level variation, we reestimate our models on the restricted sample of 14 countries with more diverse individual level communication potentials

(Model 2 and Model 4). Although, on the whole sample the communication potential does not seem to be relevant in national identity shaping (however, as we will show later in Table 5.6, the ICP is positively related with national identification in former French colonies.), the number of spoken languages in one's repertoire turns out to be influential. People who master more than two languages are more likely to identify in national terms than ethnic terms. The estimated log-odds of national identification is 0.129 higher in the case of individuals mastering more than two languages compared to those speaking only their home language. The corresponding odd-ratio is 1.138 (exp(0.129)) and it can be interpreted that being proficient in more than two languages increases the odds of national identification against ethnic identification by almost 14%. This positive effect is more striking in the last column that excludes countries with average communication potential close to one and low individual variance. On the restricted sample, the odds-ratio (exp(0.158)=1.171) indicates a 17% higher chance of identifying in national rather than ethnic terms in the case of people with more than two languages.

Our results are mostly in line with the main findings of previous works. We interpret the coefficients only in Model 3 and 4 where the language related covariate turns out to be significant. Age and religion seem to be insignificant. Female respondents tend to choose nationality over ethnicity 5.4% (the estimated odds ratio is 0.946 (exp(-0.055))) and 8.1% (the estimated odds ratio is 0.919 (exp(-0.084))) less likely compared to men in Model 3 and Model 4, respectively. The effect of gender is not clear in the empirical literature. Stronger identification in ethnic terms in the case of women is supported by Robinson (2009) (and in Warikoo (2005) in a non-African society), but no effect (Bannon et al. 2004) or positive relationship (Bossuroy 2011) is also found. One additional year in education is likely to increase the odds of national sentiment over ethnicity by 1.7% (the estimated odds ratio is 1.017= exp(0.017)) in Model 3 and 1.8% (the estimated odds ratio is 1.018= exp(0.018)) in Model 4. Living in a rural instead of urban area lowers the chance to feel more national than ethnic by about 7% (the estimated log odds are 0.934 (exp(-0.068)) and 0.925 (exp(-0.078)) in Model 3 and Model 4 respectively). The socio-economic status is significant only on the restricted sample. Since it is a composite variable based on standardized data[65] the estimated log-odds might be interpreted as it follows. A one standard deviation higher socio-economic status variable is associated with 3.5% (the estimated log odds is 1.035=exp(0.034)) higher chance of identifying in national than ethnic terms. Our findings support the modernization theory that, as we have discussed in the Section 5.2, links higher education, wealth and urbanization with higher nationalism. While most empirical analyzes arrive at this predicted positive relationship (Robinson 2009), Miles and Rochefort (1991) discuss a case study where, contrary to the conventional

[65] The data are transformed to have zero mean and unit standard deviation.

wisdom, people living in rural areas at the Nigeria-Niger border do not address more significance to their ethnicity than their nationality.

Table 5.4 The results of the multilevel analysis on the whole sample

	Model 1	Model 2	Model 3	Model 4
age	0.002 (1.36)	0.002 (1.19)	0.001 (1.22)	0.001 (1.11)
female	**-0.060*** **(-1.90)**	**-0.093***** **(-2.66)**	**-0.055*** **(-1.74)**	**-0.084**** **(-2.41)**
rural	**-0.071**** **(-2.01)**	**-0.084**** **(-2.17)**	**-0.068*** **(-1.92)**	**-0.078**** **(-2.01)**
educ	**0.019***** **(4.23)**	**0.022***** **(4.54)**	**0.017***** **(3.64)**	**0.018***** **(3.59)**
ICP	0.119 (1.34)	0.038 (0.41)	-	-
lang num=2	-	-	0.010 (0.24)	0.028 (0.59)
lang num>2	-	-	**0.129**** **(2.58)**	**0.159***** **(2.90)**
ecstat	0.013 (0.73)	**0.032*** **(1.71)**	0.013 (0.79)	**0.034*** **(1.81)**
chris	0.058 (1.35)	0.009 (0.17)	0.061 (1.42)	0.011 (0.23)
const	**-0.481***** **(-4.22)**	**-0.506***** **(-4.27)**	**-0.426***** **(-4.32)**	**-0.534***** **(-4.93)**
estimated variance of intercepts	**0.549** **(0.071)**	**0.461** **(0.069)**	**0.571** **(0.073)**	**0.468** **(0.070)**
number of individuals	19340	15768	19340	15768
number of ethnic groups	165	124	165	124
log-likelihood	-12460.168	-10177.422	-12456.15	-10171.849

Note: T-statistics are in parenthesis, 10%, 5% and 1% significance levels are labelled with *, **, and *** respectively. In the case of the estimated variance of intercepts the standard deviation of the estimated variance is in parentheses.

Our findings on the whole sample provide evidence that language learning might have positive relationship with national identification, but investigation on relevant sub-samples might bring further insight into this relationship.

As previous studies suggest (Robinson 2009), we might assume that the link between language knowledge and identity differs among former colonies since policies (including language issues as well) implemented by different European nations varied to a certain extent. The literature mainly focuses on the specialties of the rules applied by the British and the French, the main colonizers in Africa. As it is shown in Table 5.5, our sample also contains mostly British and French colonies. Although there are two former Portuguese territories (Cape Verde and Mozambique) and two territories of a

special type (Liberia and Namibia)[66] we also concentrate on the differences between the two large imperia. The French attempted to assimilate African people into the French society and the exclusive use of the French language in public affairs and education was the main tool to achieve this goal. The British did not aim to employ cultural assimilation and promoted the use of local languages in education (White 1996). While French supported centralized state administration, the British recognized the authority of traditional leaders to a greater extent (Lange 2004). Although the relationship between languages and identity might be more complicated than we first would think, based on the aforementioned well known characteristics of the European colonization practices we expect that national identity is lower in former British colonies (Robinson 2009). Data drawn from the Afrobarometer Survey presented in Table 5.5 support this assumption by showing that former French colonies experience higher nationalism (52.83%) and less variation in the country means (10.05) compared to former British colonies where the corresponding numbers are 44.62% and 16.92 respectively. In the remaining territories, the lowest level of national identification (42.78%) is associated with the highest country variation (19.74).

The interpretation of the individual standard deviations from the country means require some additional explanation. Since higher means often go hand in hand with higher standard deviations, the latter are often not interpreted in absolute terms but relative to the mean when it comes to statistical comparison. Although individual standard deviations are higher in the case of former French territories, they mean less variation relative to the country average (49.11/52.83=0.93) compared to the British colonies. The individual standard deviation in the British territories exceeds the country mean (46.97/44.62=1.053). Thus, Table 5.5 as a basic analytical tool reinforces our expectations about the higher national identification in the French colonies.

Distinguishing between former colonies refines our results (Table 5.6). None of the language related variables seem to be significant in shaping national identity on the sub-sample of British colonies (Model 1 and Model 2). The Index of Communication Potential has positive relationship with the dependent variable in the case of the French colonies (Model 3). Since the Index of Communication Potential assumes a value between 0 and 1, its estimated coefficient can be interpreted as follows. An increase of 1 percentage point (0.01) in the communication potential indicates a 0.00364 (0.364*0.01) increase in the log-odds. The corresponding odds-ratio (1.004=exp(0.00364)) suggests that a 0.01 increase in the communication potential predicts a 0.004 higher chance for identifying in national relative to ethnic terms.

[66] Liberia is considered as an independent territory. The status of Namibia is somewhat special since after the German rule starting in 1884 it was mandated to South Africa by the Leage of Nations after World War I.

Although the communication potential is not significant on the sub-sample of the remaining territories, we find that the other language variable has a significant impact (Model 5 and Model 6). Compared to those being proficient in only their home language, mastering more than two languages increases the probability of being national rather than ethnic by 52.3% (exp(0.421)=1.523).

Table 5.5 National identification in the Afrobarometer Survey countries by former colonizer

British colonies		French colonies		Other territories	
country	national identification	country	national identification	country	national identification
Botswana	39.08% (48.81)	Benin	48.71% (50.00)	Cape Verde (P)	70.67% (32.54)
Ghana	30.21% (45.94)	Burkina Faso	40.70% (49.15)	Liberia (O)	24.69% (43.14)
Kenya	47.01% (49.93)	Madagascar	67.42% (46.88)	Mozambique (P)	40.70% (49.15)
Lesotho	35.10% (47.75)	Mali	50.08% (50.02)	Namibia (O)	35.05% (47.73)
Malawi	56.97% (49.53)	Senegal	57.25% (49.49)		
Nigeria	34.13% (47.42)				
South Africa	61.41% (48.69)				
Tanzania	79.84% (40.14)				
Zambia	24.19% (42.84)				
Zimbabwe	38.25% (48.62)				
mean	44.62% (46.97)	mean	52.83% (49.11)	mean	42.78% (43.14)
standard deviation of country means	16.92	standard deviation of country means	10.05	standard deviation of country means	19.74

Note: The classification of former colonies is taken over from Bertocchi and Canova (2002). Portuguese (P), and Other (O) colonies are labeled among the other colonies. Percentages refer to the share of individuals feeling more national than ethnic in each country. Individual standard deviations from the country mean are reported in parentheses.

Table 5.6 The results of the multilevel analysis on sub-samples by former colonizer

	Model 1 (former British)	Model 2 (former British)	Model 3 (former French)	Model 4 (former French)	Model 5 (other)	Model 6 (other)
age	0.002 (1.29)	0.002 (1.21)	0.001 (0.70)	0.001 (0.60)	-0.003 (-1.02)	-0.004 (-1.17)
female	0.026 (0.64)	0.030 (0.71)	**-0.203*** (-3.46)**	**-0.205*** (-3.48)**	-0.097 (-1.16)	-0.078 (-0.92)
rural	-0.053 (-1.12)	-0.049 (-1.03)	-0.079 (-1.18)	-0.082 (-1.22)	-0.099 (-1.09)	-0.105 (-1.15)
educ	**0.011* (1.86)**	**0.010* (1.55)**	**0.031*** (3.93)**	**0.030*** (3.45)**	**0.026** (2.11)**	**0.023** (1.99)**
ICP	-0.023 (-0.19)	-	**0.364** (2.17)**	-	0.255 (1.22)	-
lang num=2	-	-0.017 (-0.30)	-	0.032 (0.42)	-	0.066 (0.52)
lang num>2	-	0.045 (0.69)	-	0.132 (1.34)	-	**0.421*** (3.05)**
ecstat	-0.035 (-1.52)	-0.035 (-1.50)	**0.057* (1.82)**	**0.055* (1.75)**	**0.106** (2.35)**	**0.107** (2.36)**
chris	0.051 (0.89)	0.052 (0.90)	0.116 (1.52)	0.119 (1.55)	0.057 (0.44)	0.066 (0.50)
const	**-0.401** (-2.49)**	**-0.417*** (-3.01)**	-0.316 (-1.64)	-0.086 (-0.54)	**-0.850*** (-3.22)**	**-0.856*** (-3.44)**
estimated variance of intercepts	**0.644 (0.110)**	**0.648 (0.110)**	**0.262 (0.076)**	**0.342 (0.093)**	**0.227 (0.081)**	**0.219 (0.079)**
number of individuals	10975	10975	5467	5467	2898	2898
number of ethnic groups	94	94	44	44	28	28
log-likelihood	-7020.3251	-7019.6575	-3617.7985	-3619.1373	-1788.5023	-1781.3084

Note: The table shows results on the total sample with 18 countries only. When we restrict our sample to the 14 countries we gain very similar results. T-statistics are in parenthesis, 10%, 5% and 1% significance levels are labeled with *, **, and *** respectively. In the case of the estimated variance of intercepts the standard deviation of the estimated variance is in parentheses.

Although the aim of this chapter is not to investigate the effects of different colonial policies on national identity in details, we intend to provide some simple explanation for our results discussed above. First, similarly to Table 5.5, Table 5.6 might be another justification to assume that British colonial rules that delegated power to local chiefs and promoted indigenous languages to a higher extent compared to the French could contribute to stronger and more persistent feelings towards the ethnic group. Second, the different patterns found on the sub-samples in Table 5.6 might be influenced by a third factor that is seemingly highly confounded with the identity of the former colonizer but not determined by it by necessity. To put it in other words, it is also possible that territories colonized by a certain European nation are similar in another aspect as well that might influence identity construction. Pre-colonial ethnic relations, relative sizes of ethnic and linguistic groups and the prestige of local languages that is

not shaped by the colonizers might influence the effect of languages on ethnic and national identity. However, it would require a separate study to explore more on this issue.

5.5 Conclusion and suggested research steps

This chapter contributes to the recent economic and political research direction that aims to reveal the main determinants of identification in national relative to ethnic terms in Sub-Saharan Africa. Similarly to previous papers, we rely on the database of the Afrobarometer Project that is designed to map out the political and democratic atmosphere in Africa. Our main contribution is that, inspired by the research results of social sciences and humanities (political studies, psycholinguistics, sociolinguistics and linguistic anthropology), we extend the possible identity shaping determinants with some language-related variables and show that language knowledge might be in positive relationship with identifying in national terms. The Index of Communication Potential (ICP) refers to the probability that a person can communicate with a randomly selected other individual within the society based on common languages. As an alternative, we control for the number of spoken languages. We estimate random intercept multilevel models with two levels on the total sample of eighteen countries and the sub-samples of former British and French colonies and other territories. The number of spoken languages performs better as an explanatory variable on the total sample and shows that people mastering more than two languages are more committed to the nation. The ICP is not significant on the total sample. However, when we distinguish between countries by former colonizer, we find that the ICP is positively related to national identification on the French sub-sample. Thus, our study serves as the first empirical evidence on a broad sample that language knowledge is in positive relationship with identity in the multilingual setting of Sub-Saharan Africa.

Our study could be developed and extended in several ways. First, as shown above, languages have different role in identity formation in British and non-British colonies. Based on relevant political and historical literature, we might identify and test the most important determinants (not necessary colonization-related) that are responsible for this difference. Moreover, our analysis does not distinguish between languages and focuses only on numbers and not the type and nature of languages. Some might expect that different tongues contribute to identity formation to various extents. European and indigenous languages might have different economic and cultural values. The size of ethnic and linguistic groups is an important determinant of networking possibilities and also influences the pressure to maintain a group identity.

This present analysis does not to incorporate additional benefits of the multilevel modeling technique. We do not include random slopes into our specification and ignore explanatory variables on the possible higher level of investigation. On the level of ethnic groups, the size and partition between countries and whether the ethnic or

linguistic group has an influential politician (president) might affect commitment to the nation. Due to the limited number of countries in the fourth round of the Afrobarometer Survey, we are not able to analyze the effect of economic development, ethnolinguistic fragmentation and the proximity of political elections on national identity. However, having the limitations of this study in mind, this work might be a promising basis for additional politically relevant research.

Appendix 5A

Table 5A.1
Variables used in the empirical analysis

Name of the variable	Name in the empirical analysis	Type of the variable	Values	Underlying database	Other remark
Dependent variable					
importance of national identity/attitudes towards ethnic identity	id_bin	binary	(0) ethnicity is at least as important as nationality (1) nationality is more important than ethnicity	Afrobarometer Survey Round 4 (Q83)	The original variable with possible values from 1 to 5 is transformed to a binary variable (see text).
Individual level variables					
age	age	continuous	above voting age	Afrobarometer Survey Round 4 (Q1)	
gender	female	categorical/ dummy	(0) male (1) female	Afrobarometer Survey Round 4 (Q101)	
living in a rural area	rural	categorical/ dummy	(0) urban (1) rural or semi-urban	Afrobarometer Survey Round 4 (URBRUR)	
years of education	educ	continuous	Minimum value is 0, maximum value is 21.	Afrobarometer Survey Round 4 (Q89) Foreign Credit Class Base Education Database	The original education categories are transformed into years of education. www.classbase.com
communication potential	ICP	continuous	between 0 and 1	Afrobarometer Survey Round 4 (Q3, Q88E) Ethnologue dictionary	http://www.ethnologue.com/ Described in more details in Appendix 5B.
number of spoken languages categories	lang num=2; lang_num>2	categorical/ dummy	the reference category is the groups of people who speak only their home language, lang=2 means that the individual speak two languages, lang>2 means that the individual speak	Afrobarometer Survey Round 4 (Q3, Q83)	

			more than 2 languages		
socio-economic status	ecstat	continuous (standardized)	one factor derived with principal component analysis	Afrobarometer Survey Round 4 (Q8A)	
religion	chris	categorical/ dummy	(0) non-christian (1) christian	Afrobarometer Survey Round 4 (Q90)	

Appendix 5B

The Index of Communication Potential (ICP)

The base of our communication potential measure is a *nxn* matrix M_k (the elements of the matrix are m_{ijk}, where *i and j=1 to n*) in each country, where *n* is the number of individuals and *k* is the country indicator (*k= 1 to 20*) in the sample. If the *i*th individual is able to communicate with the *j*th individual in country k based on spoken languages, the m_{ijk} is 1, otherwise 0. This matrix is symmetric which means that only common languages matter and other factors (willingness, ethnic conflicts) are not taken into account. And we miss the extra information of the number of commonly spoken languages. The database allows us to compute the Communication Potential at the level of individuals and countries since the sample weights are given. (Thus country level CP index is the weighted average of the individual level CP indices in a given country.)

The Index of Communication Potential (icp) of individual *i* (i=1 to N_k, where N_k is the number of individuals in country *k*) in country *k* (*k* =1 to 20) is computed as in Eq. 5B.1:

$$\mathrm{icp}_{ik} = \frac{\sum_{j=1, j\neq i}^{N_k} w_{jk} m_{ijk}}{(\sum_{j=1}^{N_k} w_{jk} - w_{ik})} = \frac{\sum_{j=1, j\neq i}^{N_k} w_{jk} m_{ijk}}{(N_k - w_{ik})}. \qquad \text{(Eq. 5B.1)}$$

where m_{ijk} assumes the value of one if individual i and j in country k can communicate, otherwise it is zero. w_{ik} and w_{jk} are the sample weights for individual *i* and *j* respectively ($\sum w_{ik} = \sum w_{jk} = N_k$).

The country level ICP (Eq. 5B.2) is computed as the weighted average of the individual indices:

$$\mathrm{ICP}_k = \frac{\sum_{i=1}^{N_k} w_{ik} \mathrm{icp}_{ik}}{N_k}. \qquad \text{(Eq. 5B.2)}$$

In this form the ICP might be interpreted as a similarity index that measures the resemblance of individuals in linguistic terms. However, it can easily be transformed to a diversity indicator. 1-ICP_k refers to the probabilty that two randomly selected people in the society cannot communicate since they have not got any common lanaguages.

As shown in Eq. 5B. 3, the above mentioned formula can be formulated as the standard fractionalization indices discussed in Ginsburg and Weber (this volume).

$$ICP_k = \frac{\sum_{i=1}^{N_k} w_{ik} icp_{ik}}{N_k} = \frac{\sum_{i=1}^{N_k} w_{ik}}{N_k} \frac{\sum_{j=1, j\neq i}^{N_k} w_{jk} m_{ijk}}{N_k - w_{ik}} = \sum_{i=1}^{N_k} \frac{w_{ik}}{N_k} \sum_{j=1, j\neq i}^{N_k} \frac{w_{jk} m_{ijk}}{N_k - w_{ik}} = \sum_{i=1}^{N_k} \sum_{j=1, j\neq i}^{N_k} p_{ik} p_{jk} m_{ijk}, \quad \text{(Eq. 5B.3)}$$

where p_{ik} ($\frac{w_{ik}}{N_k}$) and p_{jk} ($\frac{w_{jk}}{N_k - w_{ik}}$) are relative weights of individual i and j in country k respectively. Thus, we have shown that the country level ICP can be seen as a D(1,1,0) type measure in Ginsburgh and Weber (forthcoming). However, there are some minor differences. The basic units of this measure are not groups but individuals. That is why we need certain corrections in the formula to not take into account one's communication possibilities with oneself. The m_{ijk} can be interpreted as a special distance measure with possible values of one or zero. Although the traditional practice is to take groups as the basic units of diversity measurements, Bossert et al. (2011) follows the same strategy of taking individuals as the basic units in the Generalized Index of Fractionalization (GELF).

Appendix 5C

The principal component analysis (used to reveal the socio-economic status of the respondents)

Since the Afrobarometer Survey does not include questions regarding income or wealth, we apply principal component analysis to proxy the socio-economic status of individuals. The underlying questions are the following:

Q8. Over the past year, how often, if ever, have you or anyone in your family gone without:

Q8A. Enough food?

Q8B. Enough clean water from home use?

Q8C. Medicines or medical treatment?

Q8D. Enough fuel to cook your food?

Q8E. A cash income?

We derive one principal component (eigenvalue is 2.716) that accounts for 54.32% of the variance in the data and computed for 18755 observations. Since the principal component analysis is based on standardized data, summary statistics are not presented.

Appendix 5D

Table 5D.1

Ethnic groups in the sample countries

Country	Total number of listed ethnic groups in the Afrobarometer questionnaire	Number of ethnic groups included in the empirical analysis	Names of ethnic groups included in the empirical analysis	The share of included ethnic groups in the country sample
Benin	13	9	Fon	31.40%
			Adja	18.60%
			Bariba	12.37%
			Dendi	3.55%
			Yoruba	13.06%
			Otamari	7.27%
			Peulh	2.60%
			Yoa	4.33%
			Goun	6.75%
Botswana	24	13	Mokgatla	10.14%
			Mokwena	10.14%
			Mongwato	14.20%
			Mongwaketse	10.34%
			Morolong	3.77%
			Mosarwa	3.48%
			Mokalanga	16.62%
			Mokgalagadi	7.73%
			Moyei	3.29%
			Mohurutse	3.96%
			Mmirwa	4.93%
			Molete	5.60%
			Motswapong	5.80%
Burkina Faso	22	9	Mossi	59.66%
			Peul	7.10%
			Gourmatche	4.64%
			Gourounsi	6.34%
			Bobo	5.78%
			Bissa	6.44%
			Dagari	3.22%
			Samo	3.69%
			Marka	3.13%
Ghana	25	4	Akan	61.20%
			Ewe/Anglo	17.39%
			Ga/Adangbe	13.48%
			Dagomba	7.93%
Kenya	21	9	Kikuyu	21.40%
			Luo	13.89%
			Luhya	13.99%
			Kamba	11.93%
			Meru	5.66%
			Kisii	6.79%
			Kalenjin	13.17%
			Mijikenda	3.29%
			Somali	9.88%
Lesotho	31	10	Mokoena	18.74%
			Motaung	10.39%
			Mofokeng	19.29%
			Mosiea	6.49%

			Mohlakoana	10.58%
			Motsoeneng	2.78%
			Motloung	5.47%
			Motlokoa	4.17%
			Motebele	19.02%
			Lekholokoe	3.06%
Liberia	16	11	Bassa	13.41%
			Gio	7.10%
			Gola	4.71%
			Grebo	12.70%
			Kissi	4.26%
			Kpelle	25.40%
			Krahn	4.35%
			Kru	6.22%
			Lorma	8.70%
			Mano	9.41%
			Vai	3.73%
Madagascar	21	11	Antandroy	5.95%
			Antanosy	3.96%
			Antemoro	4.79%
			Antesaka	3.88%
			Betsileo	19.82%
			Betsimisaraka	14.04%
			Mahafaly	3.47%
			Merina	30.47%
			Sakalava	5.70%
			Tsimihety	3.88%
			Vezo	4.05%
Malawi	16	7	Tumbuka	11.39%
			Chewa	35.72%
			Yao	12.17%
			Ngoni	12.60%
			Lomwe	17.52%
			Manganja	5.69%
			Sena	4.92%
Mali	20	9	Bambara	33.80%
			Bobo	3.33%
			Dogon	7.62%
			Malinke	7.97%
			Peulh/Fulfulde	16.37%
			Senufo/Mianka	11.12%
			Sononke/Sarakolle	8.84%
			Sonrhai	7.97%
			Tamasheq	2.98%
Mozambique	20	10	Makua	35.09%
			Sena	10.76%
			Ndau	5.26%
			Changana	14.39%
			Chope	4.09%
			Bitonga	3.74%
			Chuabo	6.32%
			Lomue	11.11%
			Manhungie	5.73%
			Matsua	3.51%
Namibia	20	7	Wambo	56.49%
			Herero	8.72%
			Kavango	11.69%

			Afrikaaner	3.71%
			Nama	6.68%
			Damara	8.91%
			Coloured	3.80%
Nigeria	31	11	Hausa	27.23%
			Igbo	19.15%
			Yoruba	25.37%
			Fulani	4.32%
			Ibibio	4.17%
			Kanuri	2.26%
			Tiv	2.78%
			Nupe	1.60%
			Ijaw	9.06%
			Edo	2.52%
			Idoma	1.54%
Senegal	10	5	Wolof	47.46%
			Pulaar/Toucouleur	25.50%
			Serer	13.61%
			Mandinka/Bambara	7.41%
			Diola	6.03%
South Africa	14	13	English	4.91%
			Afrikaaner/Boer	15.07%
			Ndebele	1.41%
			Xhosa	15.02%
			Spedi	6.71%
			Sesotho	7.58%
			Tswana	10.70%
			Shangaan	3.65%
			Swazi	2.48%
			Venda	1.60%
			Zulu	16.97%
			Coloured	9.19%
			Indian	4.72%
Tanzania	38	8	Mchaga	6.73%
			Mpare	7.76%
			Mmakonde	7.14%
			Mnyamwezi	9.80%
			Msukuma	41.63%
			Mgogo	9.59%
			Muha	9.80%
			Mwiraqi	7.55%
Zambia	31	12	Bemba	31.08%
			Tonga	17.03%
			Lozi	10.36%
			Chewa	7.27%
			Nsenga	5.08%
			Tumbuka	6.37%
			Kaonde	3.49%
			Luvale	3.59%
			Namwanga	3.49%
			Lunda	3.98%
			Mambwe	3.09%
			Ngoni	5.18%
Zimbabwe	13	7	Ndebele	13.96%
			Shona	36.60%
			Zezuru	16.60%
			Korekore	8.21%

			Karanga	12.26%
			Manyika	7.55%
			Ndau	4.81%
Total	386	165	-	-

Note: the names of ethnic groups are taken over from the codebooks of the 4th round of the Afrobarometer Survey

References

Abdelal, R., Herrera, Y. M., Johnston, A. I., and McDermott, R., 2009. *Measuring identity: a guide for social scientists*. Cambridge: Cambridge University Press.

Abdulaziz, M. H., 1971. Tanzania's national language policy and the rise of Swahili political culture. In: W. H. Whiteley, ed. *Language use and social change*. Oxford: Oxford University Press.

Abdulaziz-Mkilifi, M. H., 1972. Triglossia and Swahili – English bilingualism in Tanzania. *Language in Society*, 1(2), 197-213.

Abrams, D. M. and Strogatz, S. H., 2003. Modelling the dynamics of language death. *Nature*, 424, 900.

Academija nauk SSSR, 1964. Atlas narodov mira. Moscow: Glavnoe Upravlenie geodezii i kartografii

Acemoglu, D., Johnson, S., and Robinson, J. A., 2001. The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5), 1369-1401.

Acemoglu, D., Johnson, S., and Robinson, J. A., 2002. Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics*, 117(4), 1231-1294.

Adams, W. M. and Anderson, D. M., 1988. Irrigation before development: Indigenous and induced change in agricultural water management in East Africa. *African Affairs*, 87(349), 519-535.

Adegbija, E., 1994. *Language attitudes in Sub-Saharan Africa*. Clevedon: Multilingual Matters.

Adegbija, E., 2007. Language policy and planning in Nigeria. *In*: R. Kaplan and R. Baldauf, eds. *Planning and Policy in Africa*. Volume 2. Clevedon: Multilingual Matters. 190-255. Also published in *Current Issues in Language Planning*, 2004, 5(3), 181-246.

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], Available from: http://www.afrobarometer.org [26.01.2015].

Aghion, P., Alesina, A., and Trebbi, F. 2004. Endogenous political institutions. *Quarterly Journal of Economics*, 119(2), 565-611.

Ahlerup, P. and Olsson, O., 2012. The roots of ethnic diversity. *Journal of Economic Growth*, 17(2), 71-102.

Ajayi, J. F. A., 1965. Christian missions in Nigeria, 1841-1891: The making of a new elite. London: Longman.

Akerlof, G. A., and Kranton, R. E., 2000. Economics and identity. *The Quarterly Journal of Economics*. 115(3), 715-753.

Akerlof, G. A., and Kranton, R. E., 2002. Identity and schooling: some lessons from the economics of education. *Journal of Economic Literature*, 40(4), 1167-1201.

Akerlof, G. A., and Kranton, R. E., 2005. Identity and the economics of organizations. *Journal of Economic Perspectives*, 19(1), 9-32.

Akerlof, G. A., and Kranton, R. E., 2010. *Identity economics: how our identities shape our work, wages, and well-being*. Princeton and Oxford: Princeton University Press.

Albaugh, E. A., 2012. *Language policies in African States – Updated, January 2012* [online]. Brunswick, Bowdoin College. Available from: http://www.bowdoin.edu/faculty/e/ealbaugh/pdf/language-policies-in-african-states-albaugh.pdf [27 Jan 2015].

Albaugh, E. A., 2014. *State-building and multilingual education in Africa*. New York: Cambridge University Press.

Albouy, D. Y., 2012. The colonial origins of comparative development: An empirical investigation: Comment. *American Economic Review*, 102(6), 3059-3076.

Alchian, A. and Demsetz, H., 1973. The property rights paradigm. *Journal of Economic History*, 33(1), 16-27.

Aldashev, A., and Danzer, A. M., 2014. Economic returns to speaking the right language(s)? Evidence from Kazakhstan's shift in state language and language of instruction. *CESIfo Working Paper No. 5068*.
Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2536267

Alderson, J. C., 2005. *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: A&C Black.

Alesina, A., Baqir, R. and Easterly, W., 1999. Public goods and ethnic divisions. *Quarterly Journal of Economics*, 114(4), 1243-1284.

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

Alesina, A. and La Ferrara, E., 2002. Who trusts others? *Journal of Public Economics*, 85(2), 207-234.

Alesina, A., and Zhukovskaya, E. 2011. Segregation and the quality of government in a cross section of countries. *American Economic Review*, 101(5), 1872-1911.

Ammon, U., 2012. Language policy in the European Union (EU). In: B. Spolsky, ed., *The Cambridge Handbook of Language Policy*, Cambridge: Cambridge University Press, 570-591.

Anderson, C. J. and Paskeviciute, A., 2006. How ethnic and linguistic heterogeneity influence the prospects for civil society: A comparative study of citizenship behavior. *The Journal of Politics*, 68(4), 783-802.

Andersson, L-G. and Janson, T., 1997. *Languages in Botswana: language ecology in Southern Africa*. Gaborone: Longman.

Angeles, L., 2007. Income inequality and colonialism. *European Economic Review*, 51(5), 1155-1176.

Angeles, L. and Neanidis, K. C., 2015. The persistent effect of colonialism on corruption. *Economica*, 82(326), 319-349.

Ansre, G., 1974. Language standardization in Sub-Saharan Africa. In: Fishman, J. A., ed., Advances in language planning. The Hague: Walter de Gruyter, 369-390.

Anyidoho, A. and Kropp Dakubu, M. E., 2008. Ghana: Indigenous languages, English, and an emerging national identity. *In*: A. Simpson, ed. *Language and national identity*, New York: Oxford University Press, 141-157.

Aparicio Fenoll, A. and Kuehn, Z., forthcoming. Does foreign language proficiency foster migration of young individuals within the European Union? *In*: B-A. Wickström and M. Gazzola, eds., *The economics of language policy*. Cambridge: MIT University Press.

Arcand, J-L. and Grin, F., 2013. Language in economic development: is English special and is linguistic fragmentation bad? *In:* E. J. Erling and P. Seargeant, eds., *English and development: policy, pedagogy and globalization*. Bristol, UK: Multilingual Matters, 243-266.

Asafo, D. R., 1997. Social class conversion: Socioeconomic status of early Christian converts in Africa. *Nordic Journal of Africa Studies*, 6(1), 81-96.

Asher, R. E. and Moseley, C., eds., 2007. *Atlas of the world's languages*. London: Routledge.

Asoni, A., 2008. Protection of property rights and growth as political equilibria. *Journal of Economic Surveys*, 22(5), 953-987.

Aspachs-Bracons O., Clots-Figueras, I., Costa-Font, J., and Masella, P., 2008. Compulsory language educational policies and identity formation. *Journal of the European Economic Association*, 6(2-3), 434-444.

Aspachs-Bracons, O., Clots-Figueras, and I., Masella, P., 2007. Identity and language policies. *Universidad Carlos III Working Paper No. 07-77(46).* Available from: http://e-archivo.uc3m.es/handle/10016/2363 [26 Jan 2015].

Atkinson, D., 1987. The mother tongue in the classroom: a neglected source? *ELT Journal*, 41(4), 241-247.

Auer, R. A., 2013. Geography, institutions, and the making of comparative development. *Journal of Economic Growth*, 18(2), 179-215.

Austin, G., 2008. The 'reversal of fortune' thesis and the compression of history: Perspectives from African and comparative economic history. *Journal of International Development*, 20(8), 996-1027.

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

Baker, C. and Jones. S. P., 1998. *Encyclopedia of bilingualism and bilingual education*. Philadelphia: Multilingual Matters.

Baldauf, R. B. and Kaplan, R. B., eds., 2004. *Language policy and planning in Africa, vol. 1, Botswana, Malawi, Mozambique and South Africa*. Clevedon: Multilingual Matters Ltd.

Baldwin, K., and Huber, J. D., 2010. Economic versus cultural differences: forms of ethnic diversity and public goods provision. *American Political Science Review*, 104(4), 644-662.

Bandyopadhyay, S. and Green, E., 2012. Pre-colonial political centralization and contemporary development in Uganda. *Afrobarometer Working paper No. 141*. Accessed from: http://www.afrobarometer.org/publications/working-papers/item/644-pre-colonial-political-centralization-and-contemporary-development-in-uganda [23 Feb 2015]

Bandyopadhyay, S. and Green, E., 2013. Nation-building and conflict in modern Africa. *World Development*, 45, 108-118.

Banfield, E. C., 1958. *The moral basis of a backward society*. New York: The Free Press.

Bannon, A., Miguel, E. and Posner, D. N., 2004. Sources of ethnic identification in Africa. *Afrobarometer Working Paper No. 44.*

Banton, M., 1997. *Ethnic and racial consciousness*, 2nd edition. New York: Longman.

Bardi, A. and Schwartz, S. H., 2003. Values and behavior: Strength and structure of relations. *Personality and Social Psychology Bulletin*, 29(10), 1207-1220.

Barrett, D. B., 1982. World Christian encyclopedia. New York: Oxford University Press.

Barrett, D. B., Kurian, G. T., and Johnson, T. M., eds., 2001. *World Christian encyclopedia: a comparative study of churches and religions in the modern world*, vol. I-II. New York: Oxford University Press.

Barro, R. J., 1991. Economic growth in a cross-section of countries. *Quarterly Journal of Economics*, 106(2), 407-443.

Barro, R. J., 1996. Democracy and growth. *Journal of Economic Growth*, 1(1), 1-27.

Bates, R., 1983. Essays on the political economy of rural Africa. Cambridge: Cambridge University Press.

Batibo, H. M., 2005. *Language decline and death in Africa: Causes, consequences and challenges*. Clevedon: Multilingual Matters.

Bayar, M., 2009. Reconsidering primordialism: an alternative approach to the study of ethnicity. *Ethnic and Racial Studies*, 32(9), 1639-1657.

Benson, C., 2010. How multilingual African contexts are pushing educational research and practice in new directions. *Language and Education*, 24(4), 323-336.

Berman, E. H., 1974. African responses to Christian mission education. *African Studies Review*, 17(3), 527-540.

Bertocchi, G., and Canova, F., 2002. Did colonization matter for growth?: An empirical exploration into the historical causes of Africa's underdevelopment. *European Economic Review*, 46(10), 1851-1871.

Betts, R. F., 2005. Assimilation and association in French colonial theory, 1890-1914. Lincoln: University of Nebraska Press.

Bjørnskov, C., 2006. Determinants of generalized trust: a cross-country comparison. *Public Choice*, 130(1-2), 1-21.

Bjørnskov, C., 2008. Social trust and fractionalization: a possible reinterpretation. *European Sociological Review*, 24(3), 271-283.

Blanton, R., Mason, D. T., and Athow, B., 2001. Colonial style and post-colonial ethnic conflict in Africa. *Journal of Peace Research*, 38(4), 473-491.

Bodenhorn, H., and Ruebeck, C. S., 2003. The economics of identity and the endogeneity of race. *NBER Working Paper No. 9962.*

Bokamba, E. G., 1991. French colonial language policies in Africa and their legacies. In: Marshall, D. F.,ed., Focus on Language Planning: Essays in honor of Joshua A. Fishman, vol. 3., Amsterdam: John Benjamins, 175-213.

Bokamba, E. G. and Tlou, J. S., 1977. The consequences of the language policies of African states vis-à-vis education. In: der Houssikian, H. and Kotey, P. A., eds., Language and linguistic problems in Africa. Columbia: Hornbeam, 35-53.

Bolt, J. and Green, E., 2015. Was the wage burden too heavy? Settler farming, profitability, and wage shares of settler agriculture in Nyasaland, c. 1900-1960. *Journal of African History*, 56(2), 1-22.

Bolt, J., and Bezemer, D., 2009. Understanding long-run African growth: Colonial institutions or colonial education? *Journal of Development Studies*, 45(1), 24-54.

Bond, M., Byrne, D., and Diamond, M. J., 1968. Effect of occupational prestige and attitude similarity on attraction as a function of assumed similarity of attitude. *Psychological Reports*, 23(3), 1167-1172.

Bossert, W., D'Ambrosio, C., and La Ferrara, E. 2011. A generalized index of fractionalization. *Economica*, 78(312), 723-750.

Bossuroy, T., 2011. Individual determinants of ethnic identification. *Dauphine Université Paris IRD, Document du Travail, DT/2011-06.*

Brown, D. S., 2000. Democracy, colonization, and human capital in Sub-Saharan Africa. *Studies in Comparative International Development*, 35(1), 20-40.

Brown, G. K. and Langer, A., 2010. Conceptualizing and measuring ethnicity. *Oxford Development Studies*, 38(4), 411-436.

Brubaker, R., Loveman, M., and Stamatov, P., 2004. Ethnicity as cognition. *Theory and Society*, 33(1), 31-64.

Bruk, S. I., and Apenchenko, V. S. (Eds.) 1964. Atlas narodov mira. Moscow: Glavnoe upravlenie geodezii i kartografii gosudarstvennogo geologicheskogo komiteta SSSR and Institut etnografii im. H. H. Miklukho-Maklaia, Akademiia nauk SSSR

Bucholtz, M., and Hall, K., 2010. Locating identity in language. In: C. Llamas and D. Watt, eds., *Language and Identities*. Edinburgh: Edinburgh University Press, 18-28.

Bunyi, G., 1999. Rethinking the place of African indigenous languages in African education. *International Journal of Educational Development*, 19(4-5), 337-350.

Bunyi, G. W., 2007. The place of African indigenous knowledge and languages in education for development: The case of Kenya. *In*: S. Nombuso Dlamini, ed. *New directions in African education*. Calgary: University of Calgary Press, 15-40.

Burton, J., Nandi, A. and Platt, L., 2010. Measuring ethnicity: Challenges and opportunities for survey research. *Ethnic and Racial Studies*, 33(8), 1332-1349.

Buss, D., 1985. Human mate selection: opposites are sometimes said to attract, but in fact we are likely to marry someone who is similar to us in almost every variable. *American Scientist*, 73, 47-51.

Buzasi, K., 2015. Languages, communication potential and generalized trust in Sub-Saharan Africa: evidence based on the Afrobarometer Survey. *Social Science Research*, 49 (1), 141-155.

Byrne, D., 1971. *The attraction paradigm*. New York: Academic Press.

Canvin, M., 2007. Language and education issues in policy and practice in Mali, West Africa. *In*: N. Rassool, K. Heugh and S. Mansoor, eds. *Global issues in languages education and development*. Clevedon: Multilingual Matters, 157-186.

Carliner, G. 1981. Wage differentials by language group and the market for language skills in Canada. *Journal of Human Resources*, 16(3), 384-399.

Cashdan, E., 2001. Ethnic diversity and its environmental determinants: Effects of climate, pathogens, and habitat diversity. *American Anthropologist*, 103(4), 968-991.

Campos, N. F. and Kuzeyev, V. S., 2007. On the dynamics of ethnic fractionalization. American *Journal of Political Science*, 51(3), 620-639.

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

Cashdan, E., 2001. Ethnic diversity and its environmental determinants: effects of climate, pathogens, and habitat diversity. *American Anthropologist, New Series*, 103(4), 968-991.

Central Statistics Office, 2005. *Report of the second national survey on literacy in Botswana.* Gaborone: The Department of Printing and Publishing Services. Available from: http://www.cso.gov.bw/templates/cso/file/File/literacy_report03.pdf [26 Jan 2015].

Central Statistical Office, 2012. *2010 Census of Population and Housing*. Volume 11: National Descriptive Tables, Lusaka: Central Statistical Office. Available from: http://catalog.ihsn.org/index.php/catalog/4124 [Accessed: 26 Jan 2015].

Chandra, K., 2006. What is ethnic identity and does it matter? *Annual Review of Political Science*, 9, 397-424.

Chee-Beng, T., 2000. Ethnic identities and national identities: some examples from Malaysia. *Identities: Global Studies in Culture and Power*, 6(4), 441-480.

Cheeseman, N., Ford, R., 2007. Ethnicity as political cleavage. Afrobarometer Conference on 'The micro-foundations of mass politics in Africa'. East Lansing, Vol. 12, No. 15.03.2007 Available from: http://www.afrobarometer.org/files/documents/working_papers/AfropaperNo83.pdf [26 Jan 2015].

Chimhundu, H., 1992. Early missionaries and the ethnolinguistic factor during the 'invention of tribalism' in Zimbabwe. *The Journal of African History*, 33(1), 87-109.

Chua, R. Y. J., Infram, P., and Morris, M. W., 2008. From the head and the heart: locating cognition-and affect-based trust in managers' professional networks. *The Academy of Management Journal*, 51(3), 436-452.

Chung, H. and Muntaner, C., 2007. Welfare state matters: a typological multilevel analysis of wealthy countries. *Health Policy*, 80(2), 328-339.

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

Church, J. and King, I., 1993. Bilingualism and network externalities. *The Canadian Journal of Economics*, 26(2), 337-345.

CIA, 2000. *The world factbook*. Washington: CIA Office Public Affairs.

Clots-Figueras, I. and Masella, P., 2013. Education, language and identity. *The Economic Journal*, 123(570), 332-357.

Cogneau, D., 2003. Colonisation, School and development in Africa. An empirical analysis. *DIAL Document de Travail 2003/I.*

Collier, P. 2000. Ethnicity, politics and economic performance. *Economics and Politics*, 12(3), 225-245.

Conklin, A. L., 1997. A mission to civilize: The republican idea of empire in France and West Africa, 1895-1930. Stanford: Stanford University Press.

Connor, W., 1972. Nation-building or nation-destroying? *World Politics*, 24(3), 319-355.

Coser, L., 1964. *Functions of social conflict*. London: Routledge and Kegan Paul.

Coulmas, F., 1992. *Language and economy*. Oxford: Blackwell.

Crafts, N., 2012. Economic history matters. *Economic History of Developing Regions*, 27(sup1), S3-S15.

Crowder, M., 1968. West Africa under colonial rule. London: Hutchinson and Co.

Crystal, D., 1998. *English as a global language*. Cambridge: Cambridge University Press.

Crystal, D., 2003. *English as a global language*. 2nd ed. Cambridge: Cambridge University Press.

Darity, W. A., Mason, P. L., and Stewart, J. B., 2006. The economics of identity: the origin and persistence of racial identity norms. *Journal of Economic Behavior and Organization*, 60(3), 283-305.

Davis, J. B., 2007. Akerlof and Kranton on identity in economics: inverting the analysis. *Cambridge Journal of Economics*, 31(3), 3494-362.

Davis, J. B., 2010. *Individuals and identity in economics*. Cambridge: Cambridge University Press.

Dearmon, J. and Grier, K., 2011. Trust and the accumulation of physical and human capital. *European Journal of Political Economy*, 27(3), 507-519.

De Groot, O. J., 2011. Culture, contiguity and conflict: on the measurement of ethnolinguistic effects in spatial spillovers. *The Journal of Development Studies*, 47(3), 436-454.

Delhey, J. and Newton, K., 2003. Who trust others?: The origins of social trust in seven societies. *European Societies*, 5(2), 93-137.

Delhey, J. and Newton, K., 2005. Predicting cross-national levels of social trust: global pattern of Nordic exceptionalism. *European Sociological Review*, 21(4), 311-327.

Desmet, K., Ortuno-Ortin, I. and Wacziarg, R., forthcoming. Linguistic cleavages and economic development. In: V. Ginsburgh, and S. Weber, eds., *Palgrave Handbook of Economics and Language*, London: Macmillan.

Desmet, K., Weber, S., and Ortuno-Ortin, I., 2009. Linguistic diversity and redistribution. *Journal of the European Economic Association*, 7(6), 1291-1318.

De Swaan, A., 1993. The evolving European language system: A theory of communication potential and language competition. *International Political Science Review*, 14(3), 241-255.

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, and 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.

Deutsch, K.W., and Foltz, W. J., 1966. *Nation-building*. New York: Atherton Press.

Diez-Roux, A. V., 2000. Multilevel analysis in public health research. *Annual Review of Public Health*, 21(1), 171-192.

Dincecco, M., Fenske, J. and Onorato, M. G., 2014. Is Africa different? Historical conflict and state development. *CSAE Working Paper WPS/2014-35* Accessed from: http://www.economics.ox.ac.uk/materials/papers/13810/csae-wps-2014-35.pdf [23 Feb 2015]

Dincer, O. C., 2011. Ethnic diversity and trust. *Contemporary Economic Policy*, 29(2), 284-293.

Doke, C. M., 1931. Report on the unification of the Shona dialects: carried out under the auspices of the Government of Southern Rhodesia and the Carnegie Corporation: presented to the Legislative Assembly, 1931, The Government of Southern Rhodesia

Dunning, T. and Harrison, L., 2010. Cross-cutting cleavages and ethnic voting: an experimental study of cousinage in Mali. *American Political Science Review*, 104(1), 21-39.

Dustmann, C., 1994. Speaking fluency, writing fluency and earnings of migrants. *Journal of Population Economics*, 7(2), 133-156.

Dyen, I., Kruskal, J. B., and Black, P., 1992. An Indo-European classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society*, 82(5), pp. iii-iv+1-132.

Easterly, W., and Levine, R., 1997. Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics*, 112(4), 1203-1250.

Easterly, W. 2001. Can institutions resolve ethnic conflict? *Economic Development and Cultural Change*, 49(4), 687-706.

Economides, N. and Himmelberg, C., 1995. Critical mass and network evolution in telecommunications. *In:* Brock, G. W., ed., *Toward a competitive telecommunication industry: selected papers from the 1994 telecommunications policy research conference*. Mahwah: Lawrence Erlbaum Associates, 47-63.

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

Egger, P., 2002. An econometric view on the estimation of gravity models and the calculation of trade potentials. *The World Economy*, 25(2), 297-312.

Emerson, R., 1961. Crucial problems involved in nation-building in Africa. *The Journal of Negro-Education*, 30(3), 193-205.

Encyclopedia Britannica, 2000. Chicago: Encyclopedia Britannica.

Englebert, P., 2000. Solving the mystery of the AFRICA dummy. *World Development*, 28(10), 1821-1835.

Englebert, P., Tarango, S., and Carter, M., 2002. Dismemberment and suffocation, a contribution to the debate on African boundaries. *Comparative Political Studies*, 35(10), 1093-1118.

Esteban, J. M. and Ray, D., 1994. On the measurement of polarization. *Econometrica*, 62(4), 819-851.

Fabian, J., 1983. Missions and the colonization of African languages: Developments in the former Belgian Congo. *Canadian Journal of African Studies*, 17(2), 165-187.

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

Fedderke, J., Luiz, J., and de Kadt, R., 2008. Using fractionalization indexes: deriving methodological principles for growth studies from time series evidence. *Social Indicators Research*, 85(2), 257-278.

Fenske, J., 2010. A causal history of Africa: Response to Hopkins. *Economic History of Developing Regions*, 25(2), 177-212.

Fenske, J., 2011. The causal history of Africa: Replies to Jerven and Hopkins. *Economic History of Developing Regions*, 26(2), 125-131.

Fenske, J., 2014. Ecology, trade, and states in pre-colonial Africa. *Journal of the European Economic Association*, 12(3), 612-640.

Fernihough, A. and O'Rourke, K. H., 2014. Coal and the European Industrial Revolution. *NBER Working Paper No. 19802*

Fidrmuc, J., Ginsburgh, V., and Weber, S., 2009. Voting on the choice of core languages in the European Union. *European Journal of Political Economy*. 25(1), 56-62.

Fieldhouse, E., Tranmer, M., and Russel, A., 2007. Something about young people or something about elections? Electoral participation of young people in Europe: evidence from a multilevel analysis of the European Social Survey. *European Journal of Political Research*, 46(6), 797-822.

Fine, B., 2009. The economics of identity and the identity of economics. *Cambridge Journal of Economics*, 33(2), 175-191.

Finseraas, H., 2008. Income inequality and demand for redistribution: a multilevel analysis of European public opinion. *Scandinavian Political Studies*, 32(1), 94-119.

Fishman, J. A., 1991. *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Clevedon: Multilingual Matters.

Fishman, J. A., 1999. *Handbook of language and ethnic identity*. New York: Oxford University Press.

Fishman, J. A., 2002. The primordialist-constructivist debate today: the language-ethnicity link in academic and in everyday-life perspective. *In:* D. Conversi, ed., *Ethnonationalism in the Contemporary World: Walker Connor and the Study of Nationalism*. London, New York: Routledge, 83-91.

Fish, M. S. and Brooks, R. S., 2004. Does diversity hurt democracy? *Journal of Democracy*, 15(1), 154-166.

Fogel, R. W., 1964, *Railroads and American Economic Growth*. Baltimore: Johns Hopkins Press.

Frankema, E., 2010. Raising revenue in the British empire, 1870-1940: How 'extractive' were colonial taxes? *Journal of Global History*, 5(3), 447-477.

Frankema, E. H. P., 2012. The origins of formal education in sub-Saharan Africa: was the British rule more benign? *European Review of Economic History*, 16(4), 335-355.

Frankema, E. and van Waijenburg, M., 2012. Structural impediments to African growth? New evidence from real wages in British Africa, 1880-1965. *The Journal of Economic History*, 72(4), 895-926.

Freitag, M., 2003. Beyond Tocqueville: the origins of social capital in Switzerland. *European Sociological Review*, 19(2), 217-232.

Freitag, M. and Bauer, P. C., 2013. Testing for measurement equivalence in surveys. Dimensions of social trust across cultural contexts. *Public Opinion Quarterly*, 77(Special issue), 24-44.

Freitag, M. and Traunmüller, R., 2009. Spheres of trust: an empirical analysis of the foundations of particularized and generalized trust. *European Journal of Political Research*, 48(6), 782-803.

Gallego, F. A., and Woodberry, R., 2010. Christian missions and education in former African colonies: How competition mattered. *Journal of African Economies*, 19(3), 294-329.

Gass, S. M. and Selinker, L., 2008. *Second language acquisition: An introductory course*. 3rd edition. New York and London: Routledge.

Garrido, S., 2011. Governing scarcity. Water markets, equity and efficiency in pre-1950s eastern Spain. *International Journal of the Commons*, 5(2), 513-534.

Gazzola, M., 2006. Managing multilingualism in the European Union: Language policy evaluation for the European Parliament. *Language Policy*, 5(4), 395-419.

Gellner, E., 1983. *Nations and nationalism*. Oxford: Blackwell.

Gennaioli, N. and Rainer, I., 2007. The modern impact of precolonial centralization in Africa. *Journal of Economic Growth*, 12(3), 185-234.

Gerring, J., Thacker, S. C., Lu, Y., Huang, W., 2015. Does diversity impair human development? A multi-level test of the diversity debit hypothesis. *World Development*, 66, 166-188.

Gerritsen, D. Lubbers, M., 2010. Unknown is unloved? Diversity and inter-population trust in Europe. *European Union Politics,* 11(2), 267-287.

Ghana Statistical Service, 2012. 2010 Population and Housing Census. Summary report of final results. Accra: Sakoa Press Limited. Available from: http://www.statsghana.gov.gh/docfiles/2010phc/Census2010_Summary_report_of_final_results.pdf [26 Jan 2015].

Gilmour, R., 2007. Missionaries, colonialism and language in nineteenth-century South Africa. *History Compass*, 5(6), 1761-1777.

Ginsburgh, V. and Weber, S., 2005. Language disenfranchisement in the European Union. *Journal of Common Market Studies*, 43(2), 273-286.

Ginsburgh, V., Ortuno-Ortin, I., and Weber, S., 2005. Disenfranchisement in linguistically diverse societies: The case of the European Union. *Journal of the European Economic Association*, 3(4), 946-965.

Ginsburgh, V. and Weber, S. forthcoming. Linguistic diversity, standardization and disenfranchisement. Measurement and consequences. *In:* B-A. Wickström and M. Gazzola, eds., *The economics of language policy*. Cambridge: MIT University Press.

Gisselquist, R. M. and McDoom, O. S., 2015. The conceptualization and measurement of ethnic and religious divisions. Categorical, temporal, and spatial dimensions with evidence from Mindanao, the Philippines. *WIDER Working Paper 2015/22*, Helsinki: United Nations University.

Githiora, C., 2008. Kenya: Language and the search for a coherent national identity. In: A. Simpson, ed. *Language and national identity*. New York: Oxford University Press, 235-251.

Graddol, D., 1997. *The future of English?* UK: The British Council.

Granovetter, M. S., 1973. The strength of weak ties. *American Journal of Sociology*, 78(6), 1360-1380.

Gray, J. P., 1999. A corrected ethnographic atlas. *World Cultures*, 10(1), 24-85.

Green, E., 2013. Explaining African ethnic diversity. *International Political Science Review*, 34(3), 235-253.

Greenberg, J. H., 1956. The measurement of linguistic diversity. *Language*, 32(1), 109-115.

Grenier, G., 1982. Language as human capital: theoretical framework and application to Spanish-speaking Americans. PhD Dissertation, Princeton: Princeton University.

Grenier, G., 1984. The effect of language characteristics on the wages of Hispanic-American males. *Journal of Human Resources*, 19(1), 35-52.

Grenoble, L. A. and Whaley, L. J., 2006. *Saving languages: An introduction to language revitalization*. Cambridge: Cambridge University Press.

Greyling, L. and Verhoef, G., 2015. Slow growth, supply shocks and structural change: The GDP of the Cape Colony in the late nineteenth century. *Economic History of Developing Regions*, 30(1), 23-43.

Grier, R. M., 1999. Colonial legacies and growth. *Public Choice*, 98(3-4), 317-335.

Grimes, J. E., and Grimes, B. F., 1996. *Ethnologue: languages of the world*. 13th edition. Dallas: SIL.

Grin, F., 1996. The economics of language: Survey, assessment, and prospects. *International Journal of the Sociology of Language*, 121(1), 17-44.

Grobler, E., Prinsloo, K. P., and van der Merwe, I. J., 1990. *Language atlas of South Africa*. Pretoria: Human Sciences Research Council.

Groves, C. P., 1964. The Planting of Christianity in Africa: 1914-1954, vol. 4., Cambridge: Lutterworth Press.

Gundelach, B., 2014. In diversity we trust: The positive effect of ethnic diversity on outgroup trust. *Political Behavior*, 36(1), 125-142.

Gustavsson, M. and Jordahl, H., 2008. Inequality and trust in Sweden: some inequalities are more harmful than others. *Journal of Public Economics*, 92(1-2), 348-365.

Habyarimana, J., Humphreys, M., Posner, D. M., and Weinstein, J. M., 2007. Why does ethnic diversity undermine public goods provisions? *American Political Science Review*, 101(4), 709-725.

Haggard, S., and Tiede, L., 2011. The rule of law and economic growth: where are we? *World Development*, 39(5), 673-685.
Hailey, L., 1945. An African Survey. Oxford: Oxford University Press.
Hale, H. E., 2004. Explaining ethnicity. *Comparative Political Studies*, 37(4), 458-485.
Hall, R. E. and Jones, C. I., 1999. Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics*, 114(1), 83-116.
Harbert, W., McConnell-Ginet, S., Miller, A., and Whitman, J., 2009. Language and poverty. Clevedon: Multilingual Matters.
Hargreaves, J. D., 1985. The making of African boundaries: Focus on West Africa. In: Asiwaju, A. I., ed., Partitioned Africans, ethnic relations across Africa's International boundaries 1884-1984. New York: St. Martin's Press.
Harris, P., 1988. The roots of ethnicity: Discourse and the politics of language construction in South-East Africa. *African Affairs*, 87(346), 25-52.
Heine, B., 1970. Status and use of African lingua francas. Afrika-Studien No. 49. Muenchen: Weltforum-Verslag GmbH.
Henderson, M. and Whately, W. C., 2014. Pacification and gender in colonial Africa: Evidence from the Ethnographic Atlas. MPRA Working Paper No. 61203. Accessed from: http://mpra.ub.uni-muenchen.de/61203/1/MPRA_paper_61203.pdf [09 March 2015]
Herfindahl, O. C., 1950. *Concentration in the steel industry*. PhD Dissertation, New York: Columbia University.
Heugh, K., 2008. Language and literacy issues in South Africa. *In*: N. Rassool, K. Heugh and S. Mansoor, eds. *Global issues in languages education and development*. Clevedon: Multilingual Matters,187-217.
Hijmans, R.J., Cameron, S.E., Parra, J. L., Jones, P. G., and Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15), 1965-1978.
Hill, C. A., 2007. The law and economics of identity. *Queen's Law Journal*, 32(2), 5-46.
Hooghe, M., Reeskens, T., Stolle, D., Trappers, A., 2009. Ethnic diversity and generalized trust in Europe. A cross-national multilevel study. *Comparative Political Studies*, 42(2), 198-223.
Hopkins, A. G., 2009. The new economic history of Africa. *The Journal of African History*, 50(2), 155-177.
Horowitz, D. L., 1985. *Ethnic groups in conflict*. Berkeley: University of California Press.
Horváth, R., 2013. Does trust promote growth? *Journal of Comparative Economics*, 41(3), 777-788.
Hox, J. J., 1995. *Applied multilevel analysis*. Amsterdam:TT-Publikaties.
Huhe, N., 2014. Understanding the multilevel foundation of social trust in rural China: evidence from the China General Social Survey. *Social Science Quarterly*, 95(2), 581-597.
Huillery, E., 2009. History matters: The long-term impact of colonial public investments in French West Africa. *American Economic Journal: Applied Economics*, 1(2), 176-215.
Hutchinson, J., 2000. Ethnicity and modern nations. *Ethnic and Racial Studies*, 23(4), 651-669.
Inglehart, R., 1999. Trust, well-being and democracy. In: M. E. Warren, ed., *Democracy and Trust*. Cambridge: Cambridge University Press, 88-120.
INSAE 2003. Troisieme Recensement General de la Population et de l'Habitation. Synthese des analyses en bref. Cotonou: Direction des Etudes Demographiques. Available from: http://www.insae-bj.org/recensement-population.html [26 Jan 2015].
Jerven, M., 2011a. The quest for the African dummy: Explaining African post-colonial economic performance revisited. *Journal of International Development*, 23(2), 288-307.
Jerven, M., 2011b. A clash of disciplines? Economists and historians approaching the African past. *Economic History of Develoing Regions*, 26(2), 111-124.
Johnson, H. B., 1967. The location of Christian missions in Africa. *Geographical Review*, 57(2), 168-202.
Johnstone, B., 2000. Locating language in identity. *In:* C. Llamas and D. Watt, eds., *Language and Identities*. Edinburgh: Edinburgh University Press, 29-36.
Jones, K., 1996. Trust as an affective attitude. *Ethics*, 107(1), 4-25.
Joseph, J. E., 2004. *Language and identity. National, Ethnic, Religious*. New York: Palgrave Macmillan.
Joshua Project, http://joshuaproject.net/ [14 March 2015]

Kahneman, D., Slovic, P., and Tversky, A., 1982. *Judgement under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press.

Kamwangamalu, N. M., 2001. The language planning situation in South Africa. *Current Issues in Language Planning*, 2(4), 361-445. Also published in R. B. Baldauf and R. B. Kaplan, eds., 2004. *Language policy and planning in Africa, vol. 1, Botswana, Malawi, Mozambique and South Africa. Clevedon*: Multilingual Matters, 197-281.

Kaplan, R. B. and Baldauf, R. B., eds., 2007. *Planning and Policy in Africa. Vol. 2. Algeria, Côte d'Ivoire, Nigeria and Tunisia.* Clevedon: Multilingual Matters.

Kaufert, J. M., 1977. Situational identity and ethnicity among Ghanaian university students. *The Journal of Modern African Studies*, 15(1), 126-135.

Kaufmann, D., Kraay, A., and Mastruzzi, M., 2004. Governance matters III: Governance indicators for 1996, 1998, 2000, and 2002. *World Bank Economic Review*, 18(2), 253-287.

Kaufman, E., 2015. Land, history or modernization? Explaining ethnic fractionalization. *Ethnic and Racial Studies*, 38(2), 193-210.

Kenya National Bureau of Statistics, 2010. *The 2009 Kenya Population and Housing Census. Volume II: Population and Household Distribution by Socio-Economic Characteristics*. Available from: http://www.knbs.or.ke/index.php?option=com_phocadownload&view=category&id=109:population-and-housing-census-2009&Itemid=599 [26 Jan 2015].

Kim, D. and Kawachi, I., 2006. A multilevel analysis of key forms of community- and individual-level social capital as predictors of self-rated health in the United States. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 83(5), 813-826.

King, K. A., and Rambow, A. C., 2012. Transnationalism, migration, and language education policy. In: B. Spolsky, ed., *The Cambridge Handbook of Language Policy*, Cambridge: Cambridge University Press, 399-417.

Knack, S. and Keefer, P., 1997. Does social capital have an economic payoff? A cross-country investigation. *The Quarterly Journal of Economics*, 112(4), 1251-1288.

Kolo, P., 2012. Measuring a new aspect of ethnicity – the appropriate diversity index. *Ibero-America Institute for Economic Research Discussion Paper No. 221*, Göttingen: Georg August University.

Konczacki, Z. A., and Konczacki, J. M., 1977. An economic history of tropical Africa, vol. 1. Abingdon: Frank Cass and Co.

Koster, F., 2013. Sociality in diverse societies: a regional analysis across European countries. *Social Indicators Research*, 111(2), 579-601.

Kayambazinthu, E., 1998. The language planning situation in Malawi. *Journal of Multilingual and Multicultural Development*, 19(5), 369-439. Also published in R. B. Baldauf and R. B. Kaplan, eds., 2004. *Language policy and planning in Africa, vol. 1, Botswana, Malawi, Mozambique and South Africa*. Clevedon: Multilingual Matters, 79-149.

Kramer, G. H., 1983. The ecological fallacy revisited: aggregate- versus individual-level findings on economics and elections, and sociotropic voting. *The American Political Science Review*, 77(1), 92-111.

Krige, D., ed., 1994. *The education atlas of South Africa*. Durban: Education Foundation.

Ladefoged, P., Glick, R. and Criper, C., 1971. *Language in Uganda*. Oxford: Oxford University Press.

Laitin, D., 1977. *Politics, language, and thought. The Somali experience*. London and Chicago: The University of Chicago Press.

Laitin, D. D., 1992. *Language repertoires and state construction in Africa*. Cambridge: Cambridge University Press.

Laitin, D., 1998. *Identity in formation. The Russian-speaking populations in the near abroad*. Ithaca and London: Cornell University Press.

Laitin, D. D., 2000. What is a language community? *American Journal of Political Science,* 44(1), 142-155.

Laitin, D. D., 2007. *Language repertoire and state construction in Africa*. Cambridge: Cambridge University Press.

Lange, M. K., 2004. British colonial legacies and political development. *World Development*, 32(6), 905-922.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A. and Vishny, R., 1999. The quality of government. *Journal of Law, Economics and Organization*, 15(1), 222-279.

Leclerc, J., (various dates) *L'Amanenagement Linguistique dans le Monde*. Available from: http://www.tlfq.ulaval.ca/AXL/afrique/afracc.htm [27 Jan 2015].

Leeson, P. T., 2005. Endogenizing fractionalization. *Journal of Institutional Economics*, 1(1), 75-98.

Levinsohn, J., 2007. Globalization and the returns to speaking English in South Africa. *In:* A. Harrison, ed., *Globalization and Poverty*, Chicago: University of Chicago Press, 629-646.

Lewis, M. P., ed., 2009. *Ethnologue: Languages of the World*. 16th ed. Dallas: SIL International. Available from: http://archive.ethnologue.com/16/ [Accessed: 26 Jan 2015].

Lewis, M. P., and Simons, G. F., 2010. Assessing Endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique*, 55(2), 103-120.

Lewis, M. P., Simons, G. F. and Fenning, C. D., eds., 2014. *Ethnologue: Languages of the world*, 17th edition, SIL International, Dallas, Texas. Online version: http://www.ethnologue.com [26 Jan 2015].

Lopes, A. J., 1998. The language situation in Mozambique. *Journal of Multilingual and Multicultural Development*, 19(5), 440-486. Also published in R. B. Baldauf and R. B. Kaplan, eds., 2004. *Language policy and planning in Africa, vol. 1, Botswana, Malawi, Mozambique and South Africa*. Clevedon: Multilingual Matters Ltd, 150-196.

Lora-Kayambazinthu, E., 2003. Language rights and the role of minority languages in national development in Malawi. *Current Issues in Language Planning*, 4(2), 146-160.

Maas, C. J. M. and Hox, J. J., 2005. Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1(3), 86-92.

MacIntyre, P. D., Noels, K. A. and Clément, R., 1997. Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, 47(2), 265-287.

Maffi, L., 2008. Biocultural diversity and sustainability. *In:* J. Pretty et al., eds., *The SAGE Handbook of Environment and Society*. London: SAGE.

Maho, J. F. 1998. Few people, many tongues: The languages of Namibia. Windhoek: Gamsberg Macmillan.

Mansour, G., 1980. The dynamics of multilingualism: The case of Senegal. *Journal of Multilingual and Multicultural Development*, 1(4), 273-293.

Mansour, G., 1993. Multilingualism and nation building, Clevedon: Multilingual Matters.

Marten, L. and Kula, N. C., 2008. Zambia: 'One Zambia, one nation, many languages'. In: A. Simpson, ed. *Language and national identity*. New York: Oxford University Press, 291-313.

Martinez-Zarzoso, I., 2003. Gravity model: An application to trade between regional blocs. *Atlantic Economic Journal*, 31(2), 174-187.

Matiki, A., 2006. Literacy, ethnolinguistic diversity and transitional bilingual education in Malawi. *The International Journal of Bilingual Education and Bilingualism*, 9(2), 239-54.

Mattes, R., 2008. The material and political bases of lived poverty in Africa. *Afrobarometer Working Papers No. 98*

Mauro, P., 1995. Corruption and Growth. *Quarterly Journal of Economics*, 110(3), 681-712.

Mavridis, D., 2015. Ethnic diversity and social capital in Indonesia. *World Development*, 67, 376-395.

McAllister, D. J., 1995. Affect- and cognition-based trust as foundations for interpersonal cooperation in organization. *The Academy of Management Journal*, 38(1), 24-59.

McLaughlin, F., 2008. Senegal: The emergence of a national lingua franca. *In*: A. Simpson, ed. *Language and national identity*. New York: Oxford University Press, 79-97.

Messick, D. M. and Mackie, D. M., 1989. Intergroup relations. *Annual Review of Psychology*, 40, 45-81.

Mesthrie, R., 2008. South Africa: The rocky road to nation building. In: A. Simpson, ed. *Language and national identity*. New York: Oxford University Press, 314-338.

Mesthrie, R., Swann, J, Deumert, A. and Leap, W. L., 2009. Introducing Sociolinguistics, 2nd edition. Edinburgh: Edinburgh University Press.

Mewes, J., 2014. Gen(d)eralized trust: women, work, and trust in strangers. European Sociological Review, 30(3), 373-386Michalopoulos, S., 2012. The origins of ethnolinguistic diversity. *American Economic Review*, 102(4), 1508-1539.

Michalopoulos, S. and Papaioannou, E., 2013. Pre-colonial ethnic institutions and contemporary African development. *Econometrica*, 81(1), 113-152.

Michalopoulos, S., and Papaioannou, E., 2015. Further evidence on the link between pre-colonial political centralization and comparative economic development in Africa. *Economics Letters*, 126(1), 57-62.

Middleton, J., 1971. Some effects of colonial rule among the Lugbara. In: Turner, V., ed., Colonialism in Africa 1870-1960. Vol. 3.: Profiles and change: African society and colonial rule. Cambridge: Cambridge University Press, 6-48.

Miles, W. F. S. and Rochefort, D. A., 1991. Nationalism versus ethnic identity in Sub-Saharan Africa. *The American Political Science Review*, 85(2), 393-403.

Molnos, A., 1969. *Language problems in Africa. A bibliographic summary (1946-67) of the present situation, with special reference to Kenya, Tanzania and Uganda*. Nairobi: East African Research Information Centre.

Montalvo, J. G., and Reynal-Querol, M., 2005. Ethnic polarization, potential conflict, and civil wars. *The American Economic Review*, 95(3), 796-816.

Montalvo, J. G. and Reynal-Querol, M., 2010. Ethnic polarization and the duration of civil wars. *Economics of Governance*, 11(2), 123-143.

Moseley, C., ed., 2010. *Atlas of the world's languages in danger*, 3rd edition, Paris: UNESCO. Available from: http://www.unesco.org/culture/en/endangeredlanguages/atlas [18 Apr 2015]

Muravyev, A., and Talavera, M., 2015. Can state language policies distort students' demand for education? *MPRA Working Paper No. 61252*. Accessed from: http://mpra.ub.uni-muenchen.de/61252/ [07 Feb 2015]

Murdock, G. P., 1959. Africa: Its people and their cultural history. New York: McGraw-Hill.

Murdock, G. P., 1967. Ethnographic Atlas. Pittsburgh: University of Pittsburgh Press.

Murdock, G. P. and White, D. R., 1969. Standard cross-cultural sample. *Ethnology*, 8(4), 329-369.

Muzale, H. R. T. and Rugemalira, J. M., 2008. Researching and documenting the languages of Tanzania. *Language Documentation and Conservation*, 2(1), 68-108.

Nagata, J. A., 1974. What is a Malay? Situational selection of ethnic identity in a plural society. *American Anthropologist*, 1(2), 331-350.

Namyalo, S., 2010. *Terminological modernization of Luganda in the field of linguistics*. Kampala: Institute of Languages Kampala, Makerere University

Namyolo, S. and Nakayiza, J., 2014. Dilemmas in implementing language rights in multilingual Uganda. *Current issues in Language Planning*, early view article, DOI: 10.1080/14664208.2014.987425

Nannestad, P., 2008. What have we learned about generalized trust, if anything? *Annual Review of Political Science*, 11, 413-436.

Nettle, D., 1999. *Linguistic diversity*. Oxford: Oxford University Press.

Nettle, D. and Romaine, S., 2000. Vanishing voices. The extinction of the world's languages. Oxford: Oxford University Press

Newton, K., 2001. Trust, social capital, civil society, and democracy. *International Political Science Review*, 22(2), 201-214.

Newton, K., 2004. Social trust: individual and cross-national approaches. *Portuguese Journal of Social Science*, 3(1), 15-35.

Nkosana, L., 2008. Attitudinal obstacles to curriculum and assessment reform. *Language Teaching Research*, 12(2), 287-312.

North, B., 2000. *The development of a common framework scale of language proficiency*. New York: Peter Lang.

North, D. C., 1990. *Institutions, institutional change and economic performance*. Cambridge: Cambridge University Press.

North, D. C., 1991. Institutions. *Journal of Economic Perspectives*, 5(1), 97-112.

North, D. C., 1992. *Transaction costs, institutions, and economic performance*. San Francisco: ICS Press.

Nunn, N., 2007. Historical legacies: A model linking Africa's past to its current underdevelopment. *Journal of Development Economics*, 83(1), 157-175.
Nunn, N., 2008. The long-term effects of Africa's slave trade. *The Quarterly Journal of Economics*, 123(1), 139-176.
Nunn, N., 2010. Religious conversion in Colonial Africa. *American Economic Review Papers and Proceedings*, 100(2), 147-152.
Nunn, N. and Puga, D., 2012. Ruggedness: The blessing of bad geography in Africa. *The Review of Economics and Statistics*, 94(1), 20-36.
Nunn, N. and Qian, N., 2011. The potato's contribution to population and urbanization: Evidence from a historical experiment. *The Quarterly Journal of Economics*, 126(2), 593-650.
Nunn, N. and Wantchekkon, L., 2011. The slave trade and the origins of mistrust in Africa. *American Economic Review*, 101(7), 3221-3252.
Nyati-Ramahabo, L., 2000. The language situation in Botswana. *Current Issues in Language Planning*, 1(2), 243-300. Also published in R. B. Baldauf and R. B. Kaplan, eds., 2004. *Language policy and planning in Africa, vol. 1, Botswana, Malawi, Mozambique and South Africa.* Clevedon: Multilingual Matters, 21-78.
Nyika, N., 2008. 'Our languages are equally important': Struggles for the revitalization of minority languages in Zimbabwe. *Southern African Linguistics and Applied Language Studies*, 26(4), 457-70.
Obondo Okoyo, T. and Sabone, I., 1986. *Twenty years of progress: An official Handbook*. Gaborone: Department of Information and Broadcasting, Publicity Section and Government Printer.
O'Campo, P., Xue, X., Wang, M. C., and Caughy, M., 1997. Neighborhood risk factors for low birthweight in Baltimore: a multilevel analysis. *American Journal of Public Health*, 87(7), 1113-1118.
Ohanessian, S. and Kashoki, M. E., 1978. *Language in Zambia*. London: International African Institute.
OIF, 2007. *Etat de la Francophonie dans le Monde 2006-2007*. Paris: Nathan.
Okediji, T. O., 2005. The dynamics of ethnic fragmentation. A proposal for an expanded measurement index. *The American Journal of Economics and Sociology*, 64(2), 637-662.
Okediji, T. O., 2011. Social fragmentation and economic growth: evidence from developing countries. *Journal of Institutional Economics*, 7(1), 77-104.
Oliver, R. 1952. The missionary factor in East Africa. London: Longmans.
Olsson, O., 2009. On the democratic legacy of colonialism. *Journal of Comparative Economics*, 37(4), 534-551.
Ortega, L., 2009. Understanding second language acquisition. Abingdon and New York: Routledge
Oshungade, I. O., 1995. The Nigerian population statistics. *In*: Ipinyomi, R. A., ed. *1995 Directory of Nigerian Statisticians*, vol. 2. 46-71.
Ostrom, E., 2000. Collective action and the evolution of social norms. *Journal of Economic Perspectives*, 14(3), 137-158.
Özcan, B. and Bjørnskov, C., 2011. Social trust and human development. *The Journal of Socio-Economics*, 40(6), 753-762.
Papagapitos, A. and Riley, R., 2009. Social trust and human capital formation. *Economic Letters*, 102(3), 158-160.
Parks, M. R. and Adelman, M. B., 1983. Communication networks and the development of romantic relationships. An expansion of uncertainty reduction theory. *Human Communication Research*, 10(1), 55-79.
Patriarca, M. and Leppänen, T., 2004. Modeling langauge competition. *Physica A*, 338(1-2), 296-299.
Patriarca, M. and Heinsalu, E., 2009. Influence of geography on language competition. *Physica A*, 388(2-3), 174-186.
Patrinos, H. A., and Velez, E., 1995. *Costs and benefits of bilingual education in Guatemala. Research report.* Washington: World Bank.
Pawlikova-Vilhanova, V., 1996. Swahili and the dilemma of Ugandan language policy. *Asian and African Studies*, 5(2), 158-170.

Paxton, P., 2007. Association membership and generalized trust: a multilevel model ascross 31 countries. *Social Forces*, 86(1), 47-76.

Pendakur, K., and Pendakur, R. 2002. Speaking in tongues: language as both human capital and ethnicity. *International Migration Review*, 36(1), 147-178.

Peters, P. E., 1994. Dividing the commons: Politics, policy, and culture in Botswana. Charlottesville: University Press of Virginia.

Peterson, D., 1997. Colonizing language? Missionaries and Gikuyu dictionaries, 1904 and 1914. *History in Africa*, 24, 257-272.

Pettigrew, T. F., 1998. Intergroup contact theory. *Annual Reviews in Psychology*, 49,65-85.

Phillipson, R., 2003. *English-only Europe? Challenging language policy*. New York: Routledge.

Pinasco, J. P. and Romanelli, L., 2006. Coexistence of languages is possible. *Physica A*, 361(1), 355-360.

Polomé, E. C. and Hill, C. P., 1980. *Language in Tanzania*. Oxford: Oxford University Press.

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), 265-283.

Pool, J., 1972. National development and language diversity. *In:* J. A. Fishman, ed., *Advances in the Sociology of Language*, vol. 2, The Hague: Mouton, 213-230.

Posner, D. N., 2004a. Measuring ethnic fractionalization in Africa. *American Journal of Political Science*, 48(4), 849-863.

Posner, D., 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.

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., Leonardi, R., and Raffaella, Y., 1994. *Making democracy work: civic traditions in modern Italy*. Princeton: Princeton University Press.

Putnam, R., 2000. *Bowling alone: the collapse and revival of American community*. New York: Simon and Schuster.

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), 137-174.

Pütz, M., ed., 1995. *Discrmination through language in Africa? Perspectives on the Namibian experience*. Berlin: Walter de Gruyter.

Quattrone, G. A. and Jones, E. E., 1980. The perception of variability within in-groups and out-groups: implications for the law of small numbers. *Journal of Personality and Social Psychology*, 38(1), 141-152.

Rabenoro, M. ed., 2013. *Langue et education: Quelle langue utiliser en classe, a Madagascar au 21eme siecle*. Cape Town: CASAS.

Rendon, S., 2007. The Catalan premium: language and employment in Catalonia. *Journal of Population Economics*, 20(3), 669-686.

Reynal-Querol, M., 2001. Ethnic and religious conflict, political systems and growth. PhD Dissertation, London: LSE.

Rice, T. and Steele, B., 2001. White ethnic diversity and community attachment in small Iowa towns. *Social Science Quarterly*, 82(2), 397-407.

Rijpma, A. and Carmichael, S., 2015. Testing Todd and matching Murdock: Global data on historical family characteristics. CGEH Working Paper No. 72, Utrecht: Utrecht University.

Rivkin, A., 1969. *Nation-building in Africa: problems and prospects*. New Brunswick: Rutgers University Press.

Robinson, W. S., 1950. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351-357.

Robinson, A. L., 2009. National versus ethnic identity in Africa: state, group and individual level correlates of national identification. *Afrobarometer Working Paper No. 112.*

Romaine, S., 2009. Biodiversity, linguistic diversity and poverty: Some global patterns and missing links. In: Harbert, W., McConnell-Ginet, S., Miller, A., and Whitman, J., 2009. Language and poverty. Clevedon: Multilingual Matters, 127-146.

Romaine, S., forthcoming**.** Language and sustainable development: Integrating the economics of language policy with poverty reduction and biodiversity conservation. *In:* B-A. Wickström and M. Gazzola, eds., *The economics of language policy*. Cambridge: MIT University Press.

Roome, 1924. Ethnographic survey of Africa, map

Rothstein, B. and Eek, D., 2009. Political corruption and social trust: an experimental approach. *Rationality and Society*, 21(1), 81-112.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C., 1998. Not so different after all: a cross-discipline view of trust. *Academy of Management Review*, 23(3), 393-404.

Rubagumya, C. M., ed., 1990. Language in education in Africa: A Tanzanian Perspective. Clevedon: Multilingual Matters.

Sachs, J. D. and Warner, A. M., 1997. Sources of slow growth in African economies. *Journal of African Economies*, 6(3), 335-376.

Salzmann, Z., Stanlaw, J., and Adachi, N., 2014. Language, culture, and society: An introduction to linguistic anthropology. Boulder: Westview Press.

Sangnier, M., 2013. Does trust favor macroeconomic stability? *Journal of Comparative Economics*, 41(3), 653-668.

Sellers, R. M., Smith, M. A., Shelton, J. N., Rowley, S. A. J., and Chavos, T. M., 1998. Multidimensional model of racial identity: a reconceptualization of African American Racial Identity. *Personality and Social Psychology Review*, 2(1), 18-39.

Shelton, J. N. and Sellers, R. M., 2000. Situational stability and variablilty in African American racial identity. *Journal of Black Psychology*, 26(1), 27-50.

Simire, G. O., 2004. Developing and promoting multilingualism in public life and society in Nigeria. In: M. J. Muthwii and A. N. Kioko, eds., *New language bearings in Africa*, Clevedon: Multilingual Matters, 135-147.

Simpson, A., ed., 2008. *Language and national identity in Africa*. Oxford: Oxford University Press.

Simpson, A. and Oyètádé, B. A., 2008. Nigeria: Ethno-linguistic competition in the giant of Africa. In: A. Simpson, ed. *Language and national identity*. New York: Oxford University Press, 172-198.

Skattum, I., 2008. Mali: In defence of cultural and linguistic pluralism. In: A. Simpson, ed. *Language and national identity*. New York: Oxford University Press, 98-121.

Skutnabb-Kangas, T. (2003). Linguistic diversity and biodiversity : the threat from killer languages. *In:* C. Mair, ed. *The politics of English as a world language : new colonial horizons in postcolonial cultural studies*. Amsterdam: Rodopi, 31-52.

Smith, A. D., 1986. *The ethnic origins of nations*. Oxford: Blackwell.

Smith, E. A., 2010. Communication and collective action: language and the evolution of human cooperation. *Evolution and Human Behavior*, 31(4), 231-245.

Snijders, T. A. B. and Bosker, R. J., 1999. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London: SAGE.

Spencer, J., 1974. Colonial language policies and their legacies in Sub-Saharan Africa. In: Fishman, J. A., ed., Advances in language planning, 163-175. The Hague: Walter de Gruyter.

Stauffer, D. and Schulze, C., 2005. Microscopic and macroscopic simulation of competition between languages. *Physics of Life Reviews*, 2(2), 89-116.

Steinhauer, H., 2005. Colonial history and language policy in Insular Southeast Asia and Madagascar. In: Adelaar, K., and Himmelmann, N., eds., The Austronesian languages of Asia and Madagascar. Abingdon: Routledge.

Stock, R., 2012. Africa South of the Sahara. A geographical interpretation, 3rd ed. New York: The Guilford Press.

Stolle, D. and Rochon, T. R., 1998. Are all associations alike?: Member diversity, assocational type, and the creation of social capital. *American Behavioral Scientist*, 42(1), 47-65.

Stolle, D., Soroka, S., and Johnston, R., 2008. When does diversity erode trust? Neighborhood diversity, interpersonal trust and the mediating effect of social interactions. *Political Studies*, 56(1), 57-75.

Stolle, D. and Harell, A., 2013. Social capital and ethno-racial diversity: learning to trust in an immigrant society. *Political Studies*, 61(1), 42-66.

Tabellini, G., 2010. Culture and institutions: Economic development and the regions of Europe. *Journal of the European Economic Association*, 8(4), 677-716.

Tait, N., ed., 1996. *A socio-economic atlas of South Africa*. Pretoria: Human Sciences Research Council.

Taylor, C. L. and Hudson, M. C., 1972. *World handbook of political and social indicators*. New Haven: Yale University Press.

Temple, J., 1998. Initial conditions, social capital and growth in Africa. *Journal of African Economies*, 7(3), 309-347.

Tequame, M. 2010. HIV, risky behavior and ethno-linguistic heterogeneity. *Center for Research in the Economics of Development Working Paper*, Namur: University of Namur.

Tokuda, Y., Fujii, S., and Inoguchi, T., 2010. Individual and country-level effects of social trust on happiness: the Asia Barometer Survey. *Journal of Applied Psychology*, 40(10), 2574-2593.

Topan, F., 2008. Tanzania: The development if Swahili as a national and official language. In: A. Simpson, ed. *Language and national identity*. New York: Oxford University Press, 252-266.

Traill, A., 1985. *Phonetic and phonological studies of !Xoo Bushman*. Hamburg: Helmut Buske Verlag

Transparency International, 2014. Corruption Perception Index 2014. Accessed from: http://www.transparency.org/cpi2014 [14 March 2015]

Traunmüller, R., 2011. Moral communities? Religion as a source of social trust in a multilevel analysis of 97 German regions. *European Sociological Review*, 27(3), 346-363.

Tsai, M-C., Laczko, L., and Bjørnskov, C., 2011. Social diversity, institutions and trust: a cross-national analysis. *Social Indicators Research*, 101(3), 305-322.

Tversky, A. and Kahneman, D., 1971. Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105-110.

Ulmer, J. T. and Johnson, B., 2004. Sentencing in context: a multilevel analysis. *Criminology*, 42(1), 137-177.

UNESCO, 2000. *World language survey: Official languages of South Africa.* Department of Arts and Culture

UNESCO, 2003. Language vitality and endangerment. International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages.

United Nations, 2014. Human Development Indicators. Accessed from: http://hdr.undp.org/en/data [14 March 2015]

USAID, 2010. *Comparative assessment of decentralization in Africa: final report and summary of findings* http://pdf.usaid.gov/pdf_docs/PNADX211.pdf [21 Dec 2014]

Uslaner, E. M., 2000-2001. Producing and consuming trust. *Political Science Quarterly*, 115(4), 569-590.

Uslaner, E. M., 2002. *The moral foundations of trust*. Cambridge: Cambridge University Press.

Uslaner, E. M. and Conley, R. S., 2003. Civic engagement and particularized trust. The ties that bind people to their ethnic communities. *American Political Research*, 31(4), 331-360.

Vaillancourt, F., 1980. *Differences in earnings by language group in Québec*. Québec: Presses de l'Université Laval.

Vaillancourt, F., 1995. Economic costs and benefits of the official languages: Some observations. In: Department of the Canadian Heritage, ed., *Official Languages and the Economy: New Canadian Perspectives*. Ottawa: Canadian Heritage Series, 103-118.

Vaillancourt, F., 1996. Language and socioeconomic status in Quebec: measurement, findings, determinants, and policy costs. *International Journal of the Sociology of Language*, 121, 69-92.

Vaillancourt, F. and Coche, O., 2009. *Official language policies at the federal level in Canada: Costs and benefits in 2006*. Vancouver: The Fraser Institute.

van der Merwe, I. J. and 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

van der Merwe, I. J. and van Niekerk, L. D., 1994. *Language in South Africa: Distribution and change*. Stellenbosch: Department of Geography, Stellenbosch University

van den Bersselaar, D., 1997. Creating 'union Ibo': Missionaries and the Igbo language. *Africa*, 67(2), 273-295.

van Parijs, P., 2011. Linguistic justice for Europe and for the world. Oxford: Oxford University Press.

Voigt, S., 2013. How (not) to measure institutions. *Journal of Institutional Economics*, 9(1), 1-26.

Wang, L. and Gordon, P., 2011. Trust and institutions: a multilevel analysis. *The Journal of Socio-Economics*, 40(5), 583-593.
Wang, C. and Steiner, B., 2015. Can ethno-linguistic diversity explain cross-country differences in social capital?: A global perspective. *Economic record*, early view, DOI: 10.1111/1475-4932.12195
Ward, I., 1940. Foreword. In: Benzies, D. Learning our language. London: Longmans.
Warikoo, N., 2005. Gender and ethnic identity among second-generation Indo-Caribbeans. *Ethnic and Racial Studies*, 28(5), 803-831.
Warren-Rothlin, A., 2009. Script choice, politics, and Bible agencies in West Africa, *Bible Translator*, 60(1), 50-66.
Welbourn, F. B. 1971. Missionary stimulus and African responses. In: Turner, V., ed., Colonialism in Africa 1870-1960. Vol. 3.: Profiles and change: African society and colonial rule. Cambridge: Cambridge University Press, 310-345.
Weldon, S. A., 2006. The institutional context of tolerance for ethnic minorities: a comparative, multilevel analysis of Western Europe. *American Journal of Political Science*, 50(2), 331-349.
Welmers, W. E., 1974. Christian missions and language policies in Africa. In. Fishman, J. A. (ed.). Advances in language planning, 191-205. The Hague: Walter de Gruyter.
White, F., 1983. The vegetation of Africa: A descriptive memoir to accompany the UNESCO/AETFAT/UNSO vegetation map of Africa. *Natural Resource Research*, 20, 1-356.
White, B., 1996. Talk about school: Education and the colonial project in French and British Africa, (1860-1969). *Comparative Education*, 32(1), 9-25.
Whiteley, W. H., 1956. The changing position of Swahili in East Africa. *Africa*, 26(4), 343-353.
Whiteley, W. H. ,1974. *Language in Kenya*. Oxford: Oxford University Press.
Wickström, B.-A., 2014. Indigenes, immigration, and integration: A welfare-economics approach to minority rights. In: F. Forte, R. Mundambi, and Navarra, P. M., eds., *A handbook of alternative theories of public economics*. Cheltenham: Edward Elgar. 227-242.
Wietzke, F.-B., 2015. Long-term consequences of colonial institutions and human capital investments: Sub-national evidence from Madagascar. *World Development*, 66, 293-307.
Williams, E., 1998. *Investigating bilingual literacy: Evidence from Malawi and Zambia*. Education Research Paper no. 24. London: British Government Department for International Development
World Bank, 2013. *GDP per capita (current USD)*. World Development Indicators. Accessed from: http://data.worldbank.org/indicator/NY.GDP.PCAP.CD [23 Feb 2015]
worldbibles.org, http://worldbibles.org [04 May 2015]
Yip, T., 2005. Sources of situational variation in ethnic identity and psychological well-being: a palm pilot study of Chinese American students. *Personality and Social Psychology Bulletin*, 31(12), 1603-1616.
Yip, T., and Fuligni, A. J., 2002. Daily variation in ethnic identity, ethnic behaviors, and psychological well-being among American adolescents of Chinese descent. *Child Development*, 73(5), 1557-1572.
Zajonc, R. B., 2001. Mere exposure: a gateway to the subliminal. *Current Directions in Psychological Science*, 10(6), 224-228.
Zak, P. J. and Knack, S., 2001. Trust and growth. *The Economic Journal*, 111(470), 295-321.
Zhang, W., and Grenier, G., 2013. How can languages be linked to economics? A survey of two strands of research. *Language Problems and Language Planning*, 37(3), 203-226.

Censuses and other surveys per country in the supplementary material of Chapter 3

Benin
Amadou Sanni, M. and Atodjinou, C. M., 2012. *Etat et dynamique des languas nationales et de la langue francaise au Benin*. Quebec: ODSEF. Available from: https://www.odsef.fss.ulaval.ca/sites/odsef.fss.ulaval.ca/files/odsef_assani_web.pdf [27 Jan 2015].

INSAE, 2003. Troisieme Recensement General de la Population et de l'Habitation. Synthese des analyses en bref. Cotonou: Direction des Etudes Demographiques. Available from: http://www.insae-bj.org/recensement-population.html [26 Jan 2015].

Botswana

Central Statistics Office, 2005. *Report of the second national survey on literacy in Botswana.* Gaborone: The Department of Printing and Publishing Services. Available from: http://www.cso.gov.bw/templates/cso/file/File/literacy_report03.pdf [26 Jan 2015].

Central Statistics Office, 2009. *Botswana Demographic Survey 2006*. Gaborone: Government Printer. Available from: http://www.cso.gov.bw/templates/cso/file/File/2006_bdsrprt.pdf [27 Jan 2015].

Burkina Faso

Institut National de la Statistique et de la Demographie, 2009a. *Recensement general de la population et de l'habitation (RGPH) de 2006. Analyse des resultats definitifs. Theme 2: Etat et structure de la population.* Ouagadougou: INSD. Available from: http://www.insd.bf/n/contenu/enquetes_recensements/rgph-bf/themes_en_demographie/Theme2-Etat_et_structure_de_la_population.pdf [27 Jan 2015].

Institut National de la Statistique et de la Demographie, 2009b. *Recensement general de la population et de l'habitation (RGPH) de 2006. Analyse des resultats definitifs. Theme 4: Education: Instruction Alphabetisation-Scolarisation.* Ouagadougou: INSD. Available from: http://www.insd.bf/n/contenu/enquetes_recensements/rgph-bf/themes_en_demographie/Theme4-Education_Instruction_Alphabetisation_Scolarisation.pdf [27 Jan 2015].

Ghana

Ghana Statistical Service, 2012. *2010 Population and Housing Census. Summary report of final results.* Accra: Sakoa Press Limited. Available from: http://www.statsghana.gov.gh/docfiles/2010phc/Census2010_Summary_report_of_final_results.pdf [26 Jan 2015].

Kenya

Kenya National Bureau of Statistics, 2010. *The 2009 Kenya Population and Housing Census. Volume II: Population and Household Distribution by Socio-Economic Characteristics.* Nairobi: Kenya National Bureau of Statistics. Available from: http://www.knbs.or.ke/index.php?option=com_phocadownload&view=category&id=109:population-and-housing-census-2009&Itemid=599 [27 Jan 2015].

Liberia

Liberia Institute of Statistics and Geo-Information Services, 2009. *2008 Population and Housing Census Final Results*. Monrovia: LISGIS. Available from: http://www.lisgis.net/pg_img/NPHC%202008%20Final%20Report.pdf [27 Jan 2015].

Liberia Institute of Statistics and Geo-Information Services, 2014. *Liberia Demographic and Health Survey 2013*. Monrovia: LISGIS. Available from: http://www.lisgis.net/pg_img/Liberia%202013%20DHS%20Final%20Report_Partial.pdf [27 Jan 2015].

Malawi

National Statistical Office, 2008. *2008 Population and Housing Census Statistical Tables on Population Characteristics (Microsoft Excel format)*. Zomba: National Statistical Office . Available

from: http://www.nsomalawi.mw/2008-population-and-housing-census/107-2008-population-and-housing-census-results.html [27 Jan 2015].

Mali
Institute National de la Statistique, 2011. *4eme Recensement General de la Population et de l'Habitation du Mali (RGPH). Resultats definitifs. Tome 1: Serie demographique.* Bamako: Institute National de la Statistique. Available from: http://www.instat-mali.org/contenu/rgph/tdemo09_rgph.pdf [27 Jan 2015].

Mozambique
Instituto Nacional de Estatistica, 2007. Recenseamento Geral da População e Habitação 2007. Maputo: Instituto Nacional de Estatistica. Available from: http://www.ine.gov.mz/operacoes-estatisticas/censos/censo-2007/rgph-2007 [27 Jan 2015].

Namibia
Namibia Statistics Agency, 2011. *Namibia 2011. Population and Housing Census Main Report*. Available from: http://www.nsa.org.na/files/downloads/Namibia%202011%20Population%20and%20Housing%20Census%20Main%20Report.pdf [27 Jan 2015].
Namibia Statistics Agency, 2003. *Namibia 2001. Population and Housing Census Main Report*. Available from: http://www.nsa.org.na/files/downloads/a5d_Namibia%202001%20Population%20and%20Housing%20Census%20Main%20Report.pdf [27 Jan 2015].

Nigeria
National Bureau of Statistics, 2010. *Report on the National Literacy Survey*. Abuja: National Bureau of Statistics. Available from: http://www.nigerianstat.gov.ng/nbslibrary/social-economic-statistics/sector-statistics [27 Jan 2015].
National Population Commission, 2009. *Demographic and Health Survey 2008*. Abuja: National Population Commission. Available from: http://www.dhsprogram.com/pubs/pdf/FR222/FR222.pdf [27 Jan 2015].

Senegal
Agence Nationale de la Statistique et de la Demographie, 2006. *Resultats definitifs du troisieme recensement general de la population et de l'habitation du Senegal (RGPH^{III}) 2002. Rapport national de presentation.* Dakar: ANSD. Available from: http://www.ansd.sn/ressources/rapports/RGPH3_RAP_NAT.pdf [27 Jan 2015].

South Africa
Statistics South Africa 2012. *Census 2011. Census in Brief*. Pretoria: Statistics South Africa, Report No. 03-01-41. Available from: http://www.statssa.gov.za/census2011/Products/Census_2011_Census_in_brief.pdf [27 Jan 2015].

Tanzania
National Bureau of Statistics, 2006. *Population and Housing Census 2002. Analytical Report Vol. X.* Dar es Salaam: National Bureau of Statistics. Available from: http://www.nbs.go.tz/nbs/takwimu/references/2002popcensus.pdf [27 Jan 2015].
National Bureau of Statistics 2014. *2012 Tanzania Basic Demographic and Socio-Economic Profile*. Literacy and Education. Dar es Salaam: National Bureau of Statistics Available from: http://www.nbs.go.tz/ [27 Jan 2015].

Uganda
Uganda Bureau of Statistics, 2006. *2002 Uganda Population and Housing Census. Analytical Report. Abridged version.* Kampala: Uganda Bureau of Statistics. Available from: http://www.ubos.org/unda/index.php/catalog/46 [27 Jan 2015].

Zambia
Central Statistical Office, 2012. 2010 *Census of Population and Housing. Volume 11: National Descriptive Tables.* Lusaka: Central Statistical Office. Available from: http://catalog.ihsn.org/index.php/catalog/4124 [26 Jan 2015].

Samenvatting

Het proefschrift heeft als doel om de relatie tussen talen en ontwikkeling in Sub-Saharisch Afrika op een vernieuwende manier te bestuderen. Hoewel taal één van de oudste en meest natuurlijk tot stand komende menselijke institutie is, is er binnen de economische en politieke wetenschap nauwelijks aandacht voor. Onderzoek naar de ontwikkelingsimpact van talen blijft beperkt tot enkele thema's en regio's. Talen worden doorgaans gezien als communicatiemiddelen die de kosten van informatiestromen, samenwerking en handel verkleinen; of als identiteitskenmerken die tot clustervorming binnen de samenleving leiden, en daarmee een rem vormen op economische activiteit en tevens conflicten veroorzaken. Andere dimensies van de taalsituatie, zoals de verspreiding van tweede talen (lingua franca's), waarvan aangenomen wordt dat ze taalbarrières verminderen, blijven onderbelicht. Onderzoek dat etnolinguïstische diversiteit overstijgt, richt zich bovendien doorgaans op ontwikkelde landen en regio's, zoals Canada, België, de Verenigde Staten en de Europese Unie.

Afrika dient als een unieke casus om de relatie tussen talen en ontwikkeling te analyseren. Ten eerste: Sub-Saharisch Afrika wordt doorgaans gezien als de meest onderontwikkelde regio ter wereld. Volgens gegevens van de Wereldbank uit 2013 over bnp per hoofd van de bevolking, zijn de vijftien armste landen ter wereld allemaal in deze regio te vinden. Onder de vijftig armste landen zijn er bovendien slechts veertien die niet in Sub-Saharisch Afrika liggen. Een groot aantal artikelen probeert dit 'mysterie van Afrika's groeitragedie' op te lossen. Ten tweede zijn Afrikaanse landen traditioneel gezien multicultureel en meertalig, en zijn de officiële talen niet inheems, maar overgenomen van de voormalige koloniale machthebbers. Het deel van de bevolking dat de officiële taal beheerst, is relatief klein. Ten derde: alhoewel een significant deel van de talen ongeschreven is of geen gestandaardiseerde orthografie heeft, blijken talen in Sub-Saharisch Afrika van groot belang; het percentage van bedreigde talen is de laagste ter wereld.

Het proefschrift heeft drie specifieke doelstellingen. Ten eerste beoogt het bij te dragen aan de literatuur over meetmethodes voor linguïstische diversiteit. Bestaande indices voor linguïstische diversiteit meten de kans dat twee willekeurig geselecteerde mensen in een samenleving verschillende primaire talen spreken. Hoewel bepaalde indicatoren erkennen dat sommige talen meer onderlinge overeenkomsten vertonen dan andere talen, kijkt geen van de meetmethodes verder dan eerste talen. Omdat de meerderheid van de bevolking in landen die taalkundig verdeeld zijn meertalig is, kan het buiten beschouwing laten van tweede talen, waarvan aangenomen wordt dat ze taalbarrières kunnen overbruggen, bij de analyse van de relatie tussen een taalsituatie en socio-economische uitkomsten tot enigszins vertekende resultaten leiden. Gebruikmakend van de onderzoeksresultaten van Afrobarometer wordt in dit proefschrift een Communicatie Potentieel Index ontwikkeld. Deze index meet de kans

dat twee willekeurig geselecteerde personen binnen een samenleving met elkaar kunnen communiceren omdat ze in ieder geval één gemeenschappelijke taal spreken.

Het tweede doel van het proefschrift is om te achterhalen of tweede talen meer maatschappelijke functies hebben dan simpelweg het faciliteren van communicatie. De belangrijkste functies van talen zijn communicatie, identiteitsvorming en het overdragen van cultuur en tradities. Als we puur naar de communicatiefunctie van talen kijken (en dus negeren dat er vaak culturele en symbolische waarden aan talen verbonden zijn), zou het niet moeten uitmaken of informatie wordt overgedragen in iemands eerste of in een andere taal: in theorie kunnen zij even efficiënt zijn (vooral in eenvoudige situaties). Echter, wanneer het gaat over de twee andere functies gaan we ervan uit dat de culturele en symbolische waarde die gehecht wordt aan eerste en tweede talen anders is. Terwijl literatuur uit de sociale en politieke wetenschappen bewijs heeft geleverd dat de identificatiefunctie en culturele waarde die aan eerste talen wordt gehecht sterk is, is over het belang van aanvullende talen in dit opzicht minder bekend.

Ten derde wil dit proefschrift ook bijdragen aan ons begrip van de wortels van de huidige taalsituatie in Sub-Saharisch Afrika. Terwijl bestaande onderzoeken zich voornamelijk richten op geografische en klimatologische determinanten van etnolinguïstische diversiteit wereldwijd, analyseert deze dissertatie hoe historische factoren invloed hebben gehad op een minder onderzochte dimensie van de taalsituatie, namelijk de taalstatus.

De bovengenoemde aspecten worden in vier hoofdstukken behandeld. Hoofdstuk 2 (gepubliceerd als CGEH Working Paper nr. 66, Universiteit Utrecht) heeft betrekking op het derde doel en onderzoekt hoe de huidige status van Afrikaanse talen bepaald is door geografische en klimatologische omstandigheden, de traditionele socio-economische karakteristieken van lokale samenlevingen en het koloniale beleid, waaronder de verspreiding en activiteiten van Christelijke missies. De status van talen wordt gemeten op een schaal die de vitaliteit van talen op vijf dimensies toetst. Talen worden gezien als minder bedreigd als ze actief gebruikt worden als eerste taal of als lingua franca, officieel erkend worden op nationaal of regionaal niveau, ononderbroken van generatie op generatie als eerste taal overgegeven worden, gedoceerd en/of gebruikt worden in het onderwijssysteem, en door de jongere generaties actief gesproken worden als eerste taal. Gegevens zijn afkomstig uit etnografische bronnen, kaarten met de locaties van missionaire gemeenschappen, online beschikbare informatie over geografische en klimatologische omstandigheden, de huidige etnolinguïstische samenstelling van samenlevingen in Sub-Saharisch Afrika, en de datum van Bijbelvertalingen. De empirische analyse is gebaseerd op een steekproef van 389 etnolinguïstische groepen die zich in 47 landen bevinden. Een belangrijke uitkomst van het onderzoek is dat pre-koloniale omstandigheden belangrijke determinanten zijn wat betreft de huidige status van talen. Lokale samenlevingen die in de negentiende eeuw al meer gecentraliseerd waren, hebben vandaag de dag een grotere kans op ontwikkelde en officieel erkende talen. De rol van

missionarisactiviteiten blijkt cruciaal: vroegere Bijbelvertalingen betekenen vaak een hogere status in het heden. De nationaliteit van de vroegere kolonisatoren heeft slechts indirect invloed via het reguleren van missionarisactiviteiten. Het indirecte effect van geografie op de status van talen blijkt belangrijker te zijn dat de directe impact. Vroege Europese contacten, gemeten door de intensiteit van vroege missionarisactiviteiten, droegen bij aan hogere ongelijkheid tussen lokale groepen, zowel in termen van taalstatus als socio-economische ontwikkeling.

Hoofdstukken 3 tot 5 zijn gebaseerd op gegevens uit het onderzoek van Afrobarometer en op de Communicatie Potentieel Index. Hoofdstuk 3 (dat gepubliceerd zal worden in *African Studies*), is een enigszins beschrijvend deel van het proefschrift, en analyseert het onderzoek van Afrobarometer als een mogelijke bron voor taalgerelateerd ontwikkelingsonderzoek, introduceert de Communicatie Potentieel Index, en presenteert beschikbare gegevens over Afrikaanse bronnen. Het laat ook zien hoe een eenvoudige grafische representatie van de Communicatie Potentieel Index gebruikt kan worden om de verschillende dimensies van de taalsituatie (zoals de etnische en taalkundige diversiteit, het communicatiepotentieel, het aantal talen in het linguïstische repertoire van de geselecteerde burger en de relatie tussen inheemse en Europese talen) te visualiseren en om de effecten van taalbeleid in kaart te brengen en te evalueren.

Hoofdstuk 4 (gepubliceerd in *Social Science Research* (Elsevier), 49, p. 141-155) en hoofdstuk 5 (te verschijnen in: M. Gazzola en B-A. Wickström, red., *The economics of language policy*, Cambridge: MIT University Press) onderzoeken de effecten van meertaligheid zoals gemeten met de Communicatie Potentieel Index en het aantal gesproken talen op twee uitvoerig onderzochte sociale fenomenen, namelijk algemeen vertrouwen en nationale identiteit, waarbij geconcludeerd wordt dat deze negatief beïnvloed worden door hoge etnolinguïstische fragmentatie. Empirisch onderzoek toont aan dat diversiteit geassocieerd wordt met een laag niveau van sociaal kapitaal en vertrouwen, zowel in ontwikkelde landen die te maken krijgen met groeiende immigratie als in ontwikkelingslanden. Hoewel individuele uitkomsten van de Communicatie Potentieel Index geen invloed lijken te hebben op algemeen vertrouwen (hoofdstuk 4), vergroot het wonen in een administratieve regio met een hogere gemiddelde uitkomst van de Communicatie Potentieel Index de kans op vertrouwen in onbekende mensen. Een ander resultaat is dat linguïstische en etnische fragmentatie positief geassocieerd worden met algemeen vertrouwen, wat de theorie ondersteunt dat als verschillende groepen intensieve interactie hebben, zij gewend raken aan diversiteit en op de lange termijn een positieve houding tegenover elkaar ontwikkelingen. Hoofdstuk 5 brengt de positieve effecten van tweede talen op een ander aspect van sociale cohesie aan het licht. Het falen van natievormend beleid in Afrikaanse staten nadat zij onafhankelijk waren geworden wordt vaak toegeschreven aan een hoge culturele en etnolinguïstische fragmentatie. Terwijl wij geen link tussen de Communicatie Potentieel Index en nationale identificatie kunnen observeren, concluderen wij wel dat mensen die meer dan twee talen spreken meer verbondenheid

voelen met de natie dan met de etnische groepen waartoe zij behoren. Hoofdstuk 4 en 5 stellen dus dat meertaligheid kan bijdragen aan sociale cohesie in de linguïstisch zo gefragmenteerde landen in Sub-Saharisch Afrika.