LINGUISTIC DIVERSITY, OFFICIAL LANGUAGE CHOICE AND NATION BUILDING: THEORY AND EVIDENCE

David D. Laitin and Rajesh Ramachandran*

July 2015

Abstract

The paper provides a theoretical framework and empirical evidence to analyze how linguistic diversity affects socio-economic development through the channel of official language choice. The problem of choosing an official language for post-colonial multilingual states is modeled as one of coordination in a society with n-linguistic groups. Through our stylized framework we highlight two factors affecting official language choice - linguistic diversity and availability of a writing tradition. It is shown that increasing linguistic diversity amplifies the problem of coordinating on the choice of an indigenous language, and increases the probability of choosing the colonial language as official. Similarly unavailability of a written indigenous language, by imposing an additional fixed cost, increases the probability of retaining the colonial language. Using both OLS and instrumental variable strategies we find strong support in the data for our theoretical framework. We explore the consequences of this unaccounted for relationship between diversity and official language choice for the cross-country empirical literature on diversity and development, and show that the negative effects attributed to diversity are mediated through the channel of language policy. Finally, we show how our theoretical framework can be usefully applied to studies on artificial states and nation building.

JEL: C7, H4, O10, P16

Keywords: Coordination Game, Language Policy, Linguistic Diversity, Nation Building.

*David D. Laitin, 423 Encina Central, Department of Political Science, Stanford University, Stanford, CA 94305. Email:dlaitin@stanford.edu. Rajesh Ramachandran, Department of Microeconomics and Management, Goethe University, Frankfurt 60323, Germany. Email:ramachandran@econ.uni-frankfurt.de

1 Introduction

One striking development of the post-world war II era has been the birth of a number of nation states, which can be classified as weak, fragile, and failing. For instance, if we were to consider non-European nation states gaining independence after 1945 as a single political entity, they would obtain an average score of 12.91 on the state fragility index constructed by Polity IV, a score corresponding to the classification "seriously" fragile. At the same time we don't fully understand which public policies can promote interethnic cooperation, increase cohesiveness and in short contribute to nation building. Besley and Persson (2010, 2011b,a) through a theoretical framework aim to understand the origins of state capacity, and show that an ineffective state is one which has made few investments in legal and fiscal capacity. The underlying roots of this ineffective state are found to lie "in the absence of common interests reinforced by non-cohesive institutions" (Besley and Persson, 2011a, pg. 395). The development of such cohesive institutions, employing the terminology of Weber (1978), is the process of state rationalization in different spheres.[1]

A large body of literature (Alesina and Ferrara 2005, Desmet et al. 2009, Easterly and Levine 1997, La Porta et al. 1999) attributes ethnolinguistic diversity to be an important factor underlying this "absence of common interests". Diversity is seen to impede provision of public goods, as well as reduce quality of government and its policies. For instance, Alesina and Ferrara (2005) note "Fragmented societies are often more prone to poor policy management and pose more politico-economic challenges than homogenous ones; it is easy to find rather voluminous evidence on this point." However as Habyarimana et al. (2007, pg. 709) note "Yet although the empirical connection between ethnic heterogeneity and the under provision of public goods is widely accepted, there is little consensus on the specific mechanisms through which this relationship operates." Thus, understanding through what channels diversity works

[1]The concept of rationalization pervades Weber's corpus. For its application to ethnicity, see Weber (1978, vol. 1, 387-95).

to impede creation of cohesive and inclusive institutions is likely to be particularly important in addressing these issues, and towards the creation of strong states.

Our paper provides a theoretical framework, and empirical evidence, to outline a heretofore unexplored channel through which linguistic diversity operates to affect socio-economic development in society. The thesis forwarded in this paper shows that linguistically diverse post-colonial states are unable to resolve problems of official language choice, and resort to retaining the colonial language, and thus do not achieve, again relying on Weber (1978), linguistic rationalization. The colonial language in turn is not the language of any indigenous group in the country and very 'distant' to the languages locally spoken. The use of a distant language, spoken by a tiny minority and hardly used for day to day interaction, imposes high costs for human capital formation, prevents effective communication across ethnic lines, and impedes public participation and discussion.[2] To analyze the implications of official language choice, we construct a weighted measure, based on Ethnologue's (Lewis et al., 2014) language trees, that calculates the average distance and exposure of the local population's languages from the official language. Empirically we find that the negative effects attributed to diversity are mediated primarily through the channel of official language choice, and accounting for this relationship in cross-country regressions renders standard diversity indices with no explanatory power.

To illustrate the link between linguistic diversity and language choice, we model the process of official language choice as one of coordination in a society with n linguistic groups. The status quo by assumption in our framework, akin to historical reality, is characterized by the use of the colonial language. Our approach rather than discriminating among the various solution concepts proposed in the literature, is rather one where we *first* employ various equilibrium selection concepts - risk dominance, Pareto dominance, equity or fairness - as well as

[2]Refer to Laitin and Ramachandran (2014) for theoretical and empirical evidence on the link between official language choice and socio-economic development. Also see Albaugh (2014) who estimates that in Sub-Saharan Africa less than 20 percent of the population on an average is able to speak the official colonial language despite more than 50 years of use as an official language.

institutional rules, such as majoritarian system, to determine the set of equilibria. Next, which is our main exercise of interest, we explore the probability (relying on any of the solution concepts) of retaining the status quo language policy as linguistic diversity in a society increases. We show that the probability of coordinating on an indigenous language is weakly decreasing in linguistic diversity. Intuitively, we assume that the cost of human capital formation increases in the distance for any individual to the official language. Thus, we should observe a decrease in human capital with an increase in the distance between the language of any two groups A and B as this distance reduces the material payoff from coordinating on the other group's language. The lower the payoff to coordination, the higher probability of the status-quo being retained.

Coordination failures for an efficient official language are more likely when we account for relative status among the linguistic groups, as the choice of any indigenous language (and the more so with language distance) will affect the economic gaps separating groups. Similarly, we show that as linguistic diversity increases, a range of decision rules for official language choice (whether the minimum quorum is unanimity or simple majority) all lead to the retention of the status quo with a "distant" language as official.

Besides diversity, availability of a well-developed written indigenous language is highlighted as an important factor affecting official language choice. In the absence of a written language, states first need to invest to create a standardized script, orthography and vocabulary before it can be used for education and administration. The cost of creating a writing tradition is modeled as imposing (i) a fixed cost; and/or (ii) uncertainty about functionality and suitability of languages that have no history of use in formal domains. It is shown that there exists fixed costs, or levels of uncertainty, such that in the absence of these a polity would choose an indigenous language, but when the cost of creating the script and orthography have to be borne they choose the colonial language. We examine and discuss language policy choices of various countries such as Angola, Indonesia, India, Rwanda, Swaziland, Tanzania, Vietnam and Zimbabwe through the lens of our model to show how our framework can be used to understand, as

well as rationalize, the observed choices.

We test our theory empirically and first show that choosing an official language not spoken by any linguistic group in the country increases the average distance from the official language. Linguistic fractionalization and the availability of a writing tradition in turn are seen to be not only significant predictors of language policy choices, but explain more than 80 percent of the variation observed in the data. Drawing from the work of Diamond (1997), who puts forth the thesis that geography was a crucial factor in determining the spread of writing traditions, we use the distance from the sites of invention of writing as an instrument for possessing a writing tradition to address concerns regarding endogeneity. The IV estimates, like the OLS estimates, provide strong support for our theoretical framework.

We next revisit the cross-country empirical literature on diversity and development in light of the theoretically and empirically demonstrated relationship between linguistic fractionalization and official language choice. In line with the existing literature, the Greenberg index of linguistic fractionalization is seen to be a negative and significant correlate of the outcomes that have been highlighted in the literature, namely, redistribution (Desmet et al. 2009, Alesina et al. 2001), quality and effectiveness of government (La Porta et al. 1999), and productivity and income levels (Alesina and Ferrara 2005, Easterly and Levine 1997), as long as we do not control for official language choice. However, once we control for the average distance from the official language, in all specifications, the coefficient on the Greenberg index of linguistic diversity becomes not only insignificant but changes sign and becomes positive. Our evidence is not meant to suggest that ELF, depending on the context, might not operate through other mechanisms highlighted in the literature such as preference, technology or strategy selection (Habyarimana et al., 2007). The aim is rather to highlight that official language choice empirically accounts for most, if not all, of the negative affects attributed directly to diversity in the cross-country literature.

The last section applies the insights from our theoretical framework to existing empirical

studies in the literature. We revisit the work by Miguel (2004) relating to nation building; and by Alesina et al. (2011) regarding artificial states. We show our framework can enrich the interpretation of their data by either concretely identifying an essential element of nation building, or helping discern the channel through which artificial states are associated with worse outcomes.

"Cohesive" (Besley and Persson, 2011a) and "inclusive" (Acemoglu and Robinson, 2012) institutions have long been recognized to be necessary building blocks for creating strong nation states, though until now little advance has been made in undercovering what determines cohesiveness or inclusiveness. Our paper makes progress in this endeavor by highlighting a specific institutional feature that constitutes or helps create such institutions. By demonstrating that one of the important channels through which diversity has harmful effects is through the choice of official language, a parameter potentially amenable to policy choices, our results also provide a basis for design of public policies that has potential to promote human capital formation, political participation, debate and development.

2 The theoretical framework

Consider a society consisting of $G \geq 1$ linguistic groups, and denote the size of any group $i \in G$ by s_i. The G groups in society are aiming to choose between the set of G indigenous and the one colonial language, $\mathscr{C}$, to act as official. Let us denote the payoff to any individual from group $i \in G$ from choosing language j to act as official by:

$$P_{ij} = \begin{cases} \Pi(1-d_{ij}) & \text{if } S_j > \kappa, \text{ where } 0 \leq d_{ij} \leq 1 \text{ and } 0 < \kappa \leq 1 \\ 0 & \text{if } S_j \leq \kappa, \ j \in G \\ \phi & \text{if } S_j \leq \kappa, \ j = \mathscr{C}, \end{cases} \tag{1}$$

where $\Pi > 0$ denotes a constant, and S_j denotes the share of the population choosing language j. The parameter d_{ij} is a measure of linguistic distance between languages i and j. It is assumed to capture the learning cost imposed in the process of human capital formation due to the official language being different from one's own language. d_{ij} is normalized and assumed to lie in the interval $[0,1]$, where 1 is the maximum possible distance between two languages i and j when they are from different language families. Thus d_{ij} will be pair specific, or in other words $d_{ij} \neq d_{ik} \; \forall \; j \neq k$.[3] Furthermore it logically follows that $d_{ij} = 0 \; \forall \; i = j$, implying all groups obtain a strictly higher payoff from their own language as compared to any other groups's language.

The above payoff formulation captures the notion of coordination, as for any language choice to have a positive payoff at least a fraction $0 < \kappa \leq 1$ of the population needs to choose the same language. Moreover, the payoff function is similar in spirit to the game of the battle-of-sexes. All groups would like to coordinate, but differ as to which language they would prefer to coordinate on. The assumption that the payoff when you fail to coordinate is equal to zero is just for simplicity and instead could be modeled as being a non-linear function of the population size, rather than in the stark way suggested by Equation 1. Concurrent with reality of post-colonial states, we also additionally assume that the *status-quo* is given by the colonial language being the official language in society. Thus in case coordination fails, the individual who remains with the colonial language is assumed to get a payoff $\phi > 0$. This is because institutional structures are already in place as far as the status-quo is concerned and hence remaining with the status-quo is more beneficial than choosing a new alternative, in case coordination fails.

The utility of an individual from any group $i \in G$ from choosing language j to act as official in turn is represented by:

$$U_{ij} = f(P_{ij}(d_{ij}, S_j), R_{ij}(P_{ij})), \tag{2}$$

[3]This is not true if both j and k are from a different language families than i, in which case $d_{ij} = d_{ik} = 1$.

where P_{ij} is the payoff given by Equation 1 and R_{ij} refers to relative ranking of group $i \in G$ resulting from the choice of language j. Thus individuals are assumed to care about not only their material payoff but also about their relative standing in society.[4] The above payoff formulation shows that there are $G+1$ pure strategy Nash equilibria in the above game. Once you are coordinating on any particular language regime, unilateral deviations are not rational. Assuming that utility is transferable and can be represented by a Utilitarian or Benthamite social welfare function will imply that we can welfare rank the $G+1$ potential pure strategy Nash equilibria.

The first key question that arises is how do countries or polities engage to decide on the official language, and second, once we determine an equilibrium selection concept or provide institutional rules for decision making, how does increasing linguistic diversity affect the probability of coordinating on an indigenous language versus a colonial one?

2.1 Decision making rules, linguistic diversity and language choice

We measure linguistic fractionalization using the index that was originally proposed by Greenberg (1956). These are a generalization of the Herfindahl index, which accounts for distance between groups, and can be interpreted as the expected distance between two randomly selected individuals in the population. The measure of linguistic fractionalization is given by:

$$LF = \sum_{i=1}^{G} \sum_{j=1}^{G} s_i s_j d_{ij}, \tag{3}$$

where s_i and s_j refer to the population shares of group $i, j \in G$ and d_{ij} refers to a measure of linguistic distance between groups i and j. It is easy to see that an increase in d_{ij} would increase the level of linguistic diversity in society.

[4] Refer to Cole et al. (1995) on how relative status/wealth concerns could be modelled as being instrumental, in the sense that individuals care about relative wealth only because final consumption is related not just to wealth, but additionally to relative wealth. The above utility function could be considered a reduced form representation of the instrumental approach.

2.1.1 Large majority populations and language choice

To see how linguistic fractionalization would affect language choice, first consider the above game as being one of pure coordination. Now assume a situation where the largest group i in society is such that $s_i > \kappa$ and $\nexists \;\; j \neq i$ s.t. $s_j > \kappa$. Given this situation it is easy to see that the unique Nash equilibrium is given by the society coordinating on the language of group i to act as official. On the other hand, you could also consider the choice of a language involving an institutionalized rule where a fraction of the population greater than κ needs to support a particular language for it become official. The above would again lead to a unique Nash equilibrium given by the choice of the indigenous language i. The above is what we consider to be representative snapshot of polities consisting of a large linguistic group - Argentina, Australia, Canada, Chile, Cambodia, Czech Republic, Laos, Slovenia, United States, Vietnam - and the country typically choosing the language of the majority linguistic group to act as official.

In the above setting an increase in linguistic fractionalization through reduction in s_i, such that $s_i^{'} < \kappa$, would imply that the choice of indigenous language i is not the unique equilibrium under both the scenarios - coordination, as well as the institutionalized rule regime. Thus, as now there is a non-zero probability that polities might remain with the status-quo, increasing linguistic diversity weakly reduces the probability of choosing an indigenous language to act as official. In what follows we analyze situations where there is no group i such that $s_i > \kappa$.

2.1.2 Risk dominance as a selection concept and linguistic diversity

In the presence of multiple equilibria one of the oft-employed solution concepts is that of risk dominance proposed by Harsanyi and Selten (1988). For simplicity assume:

ASSUMPTION 1. $d_{i\mathscr{C}} = d_{j\mathscr{C}} \geq d_{ij} \;\; \forall i, j.$

The above assumption has two implications. First, that the colonial language is equidistant from the entire set of indigenous language. Second, the distance between any two indigenous

languages is less than equal to the distance from the colonial language. This assumption is motivated by the fact that in the overwhelming majority of the cases in the data, the former colonial language belongs to a different language family compared to the language families of the set of indigenous languages.[5] As in the data when two languages are from different families they are assigned a maximum distance of 1, we make this simplifying assumption.

The above assumption implies that the coordination on any indigenous language is payoff superior to coordinating on the colonial language, or $P_{ij} \geq P_{i\mathscr{C}} \; \forall \; i \in G$. Moreover, assume that individuals have *no* relative status concerns. Under such a setup the question arises as to which equilibrium the groups will manage to coordinate on, and how does diversity affect it? To see the intuition behind it consider the game represented below in Table I, where for simplicity it is assumed that the society is made up of 2 groups A and B and $U_{ij} = f(P_{ij}(d_{ij}, S_j), R_{ij}(P_{ij})) = P_{ij}(d_{ij}, S_j)$, i.e. individuals only care about their material payoff. Recall that we have assumed that in case coordination fails, the individual who remains with the status-quo gets a payoff of $\phi > 0$. This assumption is motivated by the fact, as noted earlier, that institutional structures are already in place as far as the status-quo is concerned, and hence in case coordination fails remaining with the status-quo is more beneficial than choosing a new alternative.

Now assume individuals use risk dominance as an equilibrium selection concept. Following

Table I: **The Payoff matrix in the coordination game with two groups and three languages**

A's PARENT'S CHOICE AS MOI FOR CHILD	B's PARENT'S CHOICE AS MOI for CHILD		
	COLONIAL LANGUAGE	LANGUAGE A	LANGUAGE B
Colonial Language	$\Pi(1-d_{A\mathscr{C}}), \Pi(1-d_{B\mathscr{C}})$	$\phi, 0$	$\phi, 0$
Language A	$0, \phi$	$\Pi, \Pi(1-d_{AB})$	$0, 0$
Language B	$0, \phi$	$0, 0$	$\Pi(1-d_{AB}), \Pi$

[5] As we discuss later on, an exception is India, where Hindi and English are both from the Indo-European language family.

Harsanyi and Selten (1988) we know that the equilibrium pair (Language A, LANGUAGE A) pairwise risk dominates the equilibrium pair (Colonial Language, COLONIAL LANGUAGE) if:[6]

$$[\Pi(1-d_{A\mathscr{C}})]^2 < (\Pi-\phi)(\Pi(1-d_{AB})-\phi) \tag{4}$$

Assuming the above inequality holds and we employ the concept of risk dominance, it will imply that the groups will choose to coordinate on an indigenous language to act as official. The interesting question is then what happens to the equilibrium selected by risk dominance when linguistic diversity increases in society. As noted before, an increase in linguistic diversity corresponds to an increase in d_{AB}. Taking the derivative of the right hand side (RHS) of Equation 4, with respect to d_{AB}, shows that the RHS is strictly decreasing in d_{AB}. This implies that the probability the indigenous language equilibrium is risk dominant is decreasing in d_{AB}, or in other words increased linguistic diversity reduces the probability of coordinating on an indigenous language.

2.1.3 Incorporating equity or relative status concerns

Table I shows that coordination on the indigenous language equilibrium is Pareto dominant as $d_{i\mathscr{C}} \geq d_{ij} \ \forall \ i,j$. In this regard it might seem implausible that groups which might have possibilities to communicate will generally end up with a Pareto dominated outcome. Despite our assumption in section 2.1.2 that individuals only care about their material payoff, and that $d_{i\mathscr{C}} \geq d_{ij} \ \forall \ i,j$, a wealth of evidence from language surveys suggest that groups have strong preferences over equity or relative standing (Adegbija, 1994; Laitin, 1994; Ndamba, 2008; also refer to Heffetz and Frank (2008) for a review of the recent empirical and experimental evidence regarding preferences for social status). Also as noted before relative status concerns could also be considered to be instrumental à la Cole et al. (1995) and the above representation

[6]Observe that $d_{A\mathscr{C}} = d_{B\mathscr{C}}$ due to Assumption 1. Moreover, due to symmetry if the pair (Language A, LANGUAGE A) pairwise risk dominates the equilibrium pair (Colonial Language, COLONIAL LANGUAGE) so will the pair (Language B, LANGUAGE B) and hence transitivity will hold.

a reduced form version of the same. We now allow for relative status/wealth concerns and analyze situations where (i) individuals have relative status concerns and Assumption 1 *holds* (ii) individuals have relative status concerns and Assumption 1 is *violated.*

Relative status concerns and Assumption 1 is satisfied The assumption $d_{i\mathscr{C}} = d_{j\mathscr{C}} \geq d_{ij} \ \ \forall i, j$ implies that (i) the choice of the colonial language has the attractive feature that every groups' distance to the official language is equidistant, or in other words it makes all groups equally well (worse) off; (ii) the second inequality in turn implies that the material payoff for all groups through choosing any indigenous language is greater than equal to the payoff from the colonial language.

The above situation is typically representative of most Sub-Saharan African countries, where all indigenous languages are equidistant from the colonial language, and also the distance between any two indigenous languages is less than equal to the distance between any indigenous and the colonial language. In such a setting individuals face a tradeoff, on the one hand coordinating on any other groups' indigenous language increases their material payoff, but on the other hand it decreases their relative social standing. The utility obtained through various language policy choices is going to be determined by the relative weight given to status concerns versus the material payoff. It is important to note that we are not interested in what weights are actually accorded to the two components of utility, but given weights we are interested in answering what happens to official language choice once linguistic diversity increases.

Consider a situation where weights accorded to the two components of utility, α and β, are such that the material payoff component dominates the relative status component for a fraction of population $S_m = \bar{S} > \kappa$, when choosing some language $m \in G$. Now consider an increase in linguistic diversity such that the distance between two language groups m and n increases from d^1_{mn} to d^2_{mn}; and assume group n initially preferred the indigenous language m to the colonial language and $\bar{S} - s_n < \kappa$. Equation 2 shows that $\frac{dU_{nm}}{dd_{mn}} \leq 0$, implying that the utility is decreas-

ing in language distance, or in other words linguistic diversity. It is easy to see that there exist d^2_{mn}>d^1_{mn} such that:

$$U_{nm} = f(P_{nm}(d^1_{nm}, \bar{S}), R_{nm}(P_{nm})) > U_{n\mathscr{C}} = f(P_{n\mathscr{C}}(d_{\mathscr{C}n}, 1-\bar{S}-s_n), R_{n\mathscr{C}}(P_{n\mathscr{C}})) =$$
$$U_{n\mathscr{C}} = f(P_{n\mathscr{C}}(d_{\mathscr{C}n}, 1-\bar{S}-s_n), R_{n\mathscr{C}}(P_{n\mathscr{C}})) > U_{nm} = f(P_{nm}(d^2_{nm}, \bar{S}), R_{nm}(P_{nm})) \quad (5)$$

Thus, we again see that increasing linguistic diversity would imply that groups tend to stick with the status-quo more often.

The above analysis suggests that if there were to exist a language e such that $d_{i\mathscr{C}} > d_{ie} = d_{je} \ \forall i, j \in G$, then such a language choice would make all groups unambiguously better off compared to the status-quo, and we should see even linguistically diverse polities moving to the Pareto dominant equilibrium. Does this prediction seem to be borne out in reality? An interesting example supporting the above prediction is the case of Indonesia. Indonesia is highly linguistically diverse with a number of ethnic groups, speaking an estimated 600 languages (Paauw, 2009). Javanese is the language of the largest linguistic group, comprising about 45 percent of the population, and had been the primary language of politics and economics, and the language of courtly, religious, and literary tradition, making it seemingly the obvious choice to act as the official language at independence (De Swaan, 2013). Interestingly enough we observe that Indonesia actually chose Bahasa Indonesian as the official language. Bahasa Indonesian is a standardized register of Malay, an Austronesian language that has been used as a lingua franca in the Indonesian archipelago for centuries. The underlying reasons behind this choice can be rationalized through the lens of our framework, and is also strongly supported by historical evidence. The use of a lingua franca widely spoken and understood by a vast majority of the population meant an unambiguous decline in the language distance, increasing the material payoff $P_{ie} \ \forall i \in G$. Secondly, as the language was not the language of any sizeable ethnic group in the country, the choice of this neutral language meant that relative status concerns or R_{ie} were

not (or minimally) affected. This implies that there exists a Pareto dominant equilibrium for all groups concerned to coordinate on.

In line with our theoretical hypothesis, Paauw (2009, 2) discusses how the need to avoid resentment and fears by other ethnic groups regarding domination by Javanese in political and economic domains, if Javanese was chosen as official, was one of the principal reasons underlying the choice of Indonesian. As Errington (1998, 51) adds the "very un-nativeness [of Malay] has been the key to the success of Indonesian language development." This said it should be mentioned that there were several other contributing factors whose role cannot be minimized. For instance, Anderson (1990) discusses the role of the Javanese elite and how the willingness to accept Indonesian was a magnanimous concession on their part.[7] Another key event is the 1942 Japanese occupation of Indonesia, which has been referred to as one of the most decisive moments in the development of Indonesian (Alisjahbana, 1962). Vickers (2013) discusses the Japanese role in the economic, political and social dismantling of the Dutch colonial service. They importantly forbade the use of Dutch for any purpose, resorted to using Indonesian as the main language of administration and public affairs, with the ultimate aim of replacing it with Japanese. This meant that with the defeat of Japan, Dutch-speaking elites were unable to benefit from their linguistic capital through the re-introduction of Dutch as the official language of their newly independent state. A final reason is that the importance of Dutch as an international language was much more limited than English or French, making it easier to dispel with Dutch.[8]

Another interesting example is the case of Tanzania, which is the only post-colonial state in Sub-Saharan Africa offering the entire span of *primary* schooling in a non-colonial language, namely, Swahili. Swahili, a language spoken by the natives of the coastal mainland spread to the rest of the Swahili coast starting the 2^{nd} century AD, initially as a fisherman's language, and

[7] Although it should be mentioned that he also points out that this a sentiment exhibited mainly by the Javanese of future generations.

[8] Also refer to Dardjowidjojo (1998) who discusses the fear of domination by the Malays and Tagalogs, the majority linguistic groups in Malaysia and Philippines, as one of the important reasons why English was given an important role post-independence in both contexts.

eventually as the language of trade and commerce. The fact that Swahili was not identified with a specific ethnic group or social class implied it could be easily accepted as a politically neutral alternative by all groups in Tanzania (European Commission, Directorate General for Translation, 2011). Here too it should be pointed out that there were other historical factors which led to more intensive promotion of Swahili in Tanzania as compared to neighboring Kenya, where it too had served the role of a lingua franca. One important factor was that the Germans during their occupation of Tanzania from 1886 to 1918 designated Swahili as a colony-wide official administrative language, whereas the British in Kenya did not do so. Another crucial factor was the role of Julius Nyerere, Tanzania's first president, who promoted Ujamaa, a nationalist and pan-Africanist ideology that revolved around reliance on Swahili instead of on European languages.

Relative status concerns and Assumption 1 is not satisfied Consider a society made of two groups A and B; assume that $d_{A\mathscr{C}} < d_{B\mathscr{C}}$ and $d_{AB} < d_{B\mathscr{C}}$ and $\kappa = 1$. The above implies that the distance to the colonial language for group A is less than that for group B, but however for group B the distance from the language of group A is less than the distance from the colonial language. Assume that weights on the material payoff and relative status are such that $U_{BA} > U_{B\mathscr{C}}$ is satisfied. In other words the increased payoff from choosing language A for group B overweighs the loss arising from the decline in relative status. Under such a scenario it is plausible that groups are able to communicate, bargain and coordinate on language A. Now assume that the distance between the two groups increase such that $d_{AB} = d_{B\mathscr{C}}$. This would imply that a switch to language A does not increase the material payoff component but reduces the relative standing for group B, in other words $U_{BA} < U_{B\mathscr{C}}$. In this case group B would not be willing to move, and as $\kappa = 1$ the country would remain with the status-quo. We thus again observe how linguistic diversity reduces the probability of choosing an indigenous language.

Are there any real world examples that seem to follow the pattern suggested above? The

case of India indeed closely parallels the situation described above. India is comprised of a multitude of languages, where in the Northern part of India the languages come from the Indo-European family, with the Hindi speakers comprising around 40 percent of the population. On the other hand, in South India, the languages come from the language family called the Dravidian. In the language of our model, if we were to consider only North India, and assume group A to be Hindi speakers, all group B languages also come from the Indo-European family, and hence pertain to the setting where $d_{A\mathscr{C}} \leq d_{B\mathscr{C}}$ and $d_{AB} < d_{B\mathscr{C}}$. However once we consider the Southern states the situation resembles the case where $d_{AB} = d_{B\mathscr{C}}$, as now group B speakers come from the Dravidian family, or in other words are more distant to language A.[9] Our framework suggests that as Tamil (or Dravidian) language speakers had nothing to gain by switching to Hindi but face a loss in the relative status, they would be strongly opposed to making Hindi the official language.

History reveals exactly the same dynamics as suggested by our framework. The India National Congress was keen to institute Hindi as the official language of India, with as early as 1918 Mahatma Gandhi establishing the *Dakshin Hindi Prachar Sabha* (Institute for the Propagation of Hindi in South India). In 1937 the Indian National Congress won the elections in Madras Presidency, with Rajaji becoming the chief minister. Rajaji was an ardent supporter of promoting Hindi in South India and announced his intention to introduce Hindi language teaching in secondary schools by issuing a policy statement to this effect (More, 1997). This announcement set the stage for the first anti-Hindi agitations to break out in Tamil Nadu in particular, and in South India in general. The agitation was marked by fasts, protest marches, processions, picketing of schools teaching Hindi and government offices, anti-Hindi conferences, observing an anti-Hindi day and black flag demonstrations (Irschick, 1986; Ramaswamy, 1997). It is instructive to note that the opposition primarily came from the more distant Dra-

[9] To fix ideas you could assume that in the setting of only North India, group B are Gujarati speakers, whereas when we include South India, consider group B to be Tamil speakers (as Tamil is equidistant from both Hindi and English this would imply $d_{Hindi-Tamil} = d_{English-Tamil}$).

vidian speaking language groups, and not the other non-Hindi Indo-European languages, as suggested by our framework. With the outbreak of the Second World War, the Congress government resigned to protest India's participation in it and the compulsory Hindi order was rescinded in 1940. The language issue again came to the fore at independence in 1947 and the process of drafting the constitution. The members of the Hindi speaking provinces argued for adopting Hindi as the sole official language and moved a number of pro-Hindi amendments (Austin, 1999). These were strongly resisted by the anti-Hindi block which favored retaining English as the official language (Annamalai, 1979). After three years of debate a compromise was reached where for the next fifteen years, both English and Hindi, would be the languages of the Indian Union. The announcement that the situation could be revisited meant that the fears of the Dravidian language speakers were not quelled, and eventually resulted in the introduction of the official language act of 1963 by Nehru. The proposed bill was meant to remove the restriction which had been placed by the Constitution on the use of English after a certain date, i.e. 1965. The bill was hotly debated with Annadurai, a leading Hindi opponent from Tamil Nadu, pleading for an indefinite continuation of the status-quo and argued that continued use of English as the official language would "distribute advantages or disadvantages evenly" among Hindi and non-Hindi speakers (Ramachandran, 1975, 65). The situation was finally resolved when in 1967 Indira Gandhi passed an amendment and guaranteed the "virtual indefinite policy of bilingualism" (Chandra, 2000). The above historical example nicely illustrates the role of relative status concerns and linguistic diversity highlighted in the framework.

2.2 The role of writing tradition - Incorporating fixed costs and uncertainty

The previous discussion highlights how increasing linguistic diversity, as measured by the linguistic distance between two groups, both reduces the material payoff and relative status ranking, and in turn the utility of coordinating on either of the group's language, and makes the

probability of retaining the status-quo higher. However, when we examine language policy choices in the real world, we observe that countries such as Estonia, Georgia, India and Iran, have all chosen an indigenous language to act as (co-)official, whereas states with much lower levels of diversity such as Angola, Mozambique and Zimbabwe, have exclusively retained the colonial language to act as official. What can explain this discrepancy between our theory and the observed outcomes in the real world? We now highlight a second important factor - *the availability of a developed writing script for a major linguistic groups' language in the country* - affecting language policy choices in post-colonial states. The rationale behind why availability of a written script should affect official language choice is straightforward. In the absence of a written indigenous language, the process of creation of a standardized script, orthography, and vocabulary to deal with modern scientific concepts has to be undertaken before any indigenous language can be installed as official.

In light of the theoretical framework presented before, lack of availability of a standardized writing script can be understood as affecting language policy choices through either or both of the following channels: (i) imposing a fixed cost for creation of a standardized script, orthography, and vocabulary; (ii) uncertainty associated with suitability of and returns to a language that has never been employed in formal domains.[10] To see how this would affect the process of language choice, consider a society of two groups A and B and $\kappa = 1$. Moreover, assume that U_{BA}>$U_{B\mathscr{C}}$, implying both groups prefer language A to the colonial language, and allowing for communication that facilitates coordination will imply that language A is chosen. Now assume that they need to invest a fixed amount denoted by φ to standardize language A. Thus their

[10] Another important political economy mechanism is emphasized by Laitin (2000, 2004); to get a new written language, you need to rely on the civil service which has an interest in maintaining the colonial status quo, and will therefore raise the costs of vocabulary development through shirking.

material payoff can now be represented by:

$$P_{ij} = \begin{cases} \Pi(1-d_{ij}) - \varphi & \text{if } S_j > \kappa, \text{where } 0 \leq d_{ij} \leq 1 \text{ and } 0 < \kappa \leq 1 \\ -\varphi & \text{if } S_j \leq \kappa, \end{cases} \tag{6}$$

It is easy that there exist levels of fixed cost φ such that in the absence of it $U_{BA} > U_{B\mathscr{C}}$, whereas in the presence of it $U_{BA} < U_{i\mathscr{C}}$; implying for a given level of linguistic diversity states with a writing tradition would have chosen the indigenous language whereas in the presence of these fixed costs they prefer to remain with the status-quo.

An alternative way to capture how absence of writing tradition affects language choice is through the notion of uncertainty associated with technological choices that have never been utilized before. Given that most Sub-Saharan African states had oral traditions and have no experience in utilizing their languages in formal domains could mean that individuals are uncertain about their suitability for use in formal domains, or erroneously believe that these oral languages are unsuitable for modern scientific communication.[11] Moreover, a policy of deliberate denigration of local languages in favor of the imperial languages by the colonialists has led to low status of indigenous languages, even among their native speakers. As Adegbija (1994, 22) notes "the neglect suffered by these languages and the fact that they were not used in things that mattered and counted in the national plane, naturally built and institutionalized negative attitudes around them, especially in official domains. Such attitudes have been difficult to remove after independence." Thus akin to the fixed cost channel allowing for returns from coordinating on an indigenous language to vary over the interval $[\underline{X}, \Pi]$ due to uncertainty, where $\underline{X} < \Pi$, would again imply that there exist a $\underline{X}$ such that a country for a given level of linguistic diversity would coordinate on an indigenous language in absence of this uncertainty but remains with the status-quo in the presence of it.

[11] Refer to Bourdieu (1991) for a critique of the position that African languages are inherently unsuitable for science.

A final point to note is that our model also does well in predicting choices of linguistically homogenous states but without a written tradition, namely, states such as Botswana, Burundi, Rwanda, Somalia and Swaziland. Our model on the one hand would suggest that as these states are largely homogenous, according to the analysis presented in section 2.1.1, they should choose the majority group language to act as official. On the other hand, the lack of a written tradition through imposition of fixed costs and uncertainty associated with their suitability and returns should reduce the probability of choosing an indigenous language. The reality seems to tailor well with the predictions of the model; Botswana, Burundi, Rwanda, Somalia and Swaziland all have chosen to institute the language of the majority group as official, however their de facto role in society remains severely restricted. In most cases the official indigenous language is not even used for the entire span of primary schooling, and the knowledge of the former colonial language remains indispensable in order to obtain higher education and consequently socioeconomic mobility. Thus, truly overcoming the constraints of linguistic diversity in the choice of an official language seems to require a writing tradition.

3 Empirical evidence for the theoretical framework

3.1 Why do we care about official language choice and creating a measure of distance from official language

The theoretical framework shows that increasing linguistic diversity and the absence of a written tradition results in increasing the probability of retaining the status quo, i.e. the colonial language as official. The colonial language in turn is characterized by being "distant" to the languages spoken locally, and consequently increases distance to the official language. In this regard two important questions arise: (1) how do we operationalize the notion of distance between languages? (2) Why do we care about distance from the official language?

To measure the distance between languages of the indigenous groups in a country and their official language, we use the measure based on Ethnologue's linguistic tree diagrams. The distance between any two languages i and j based on Fearon (2003) is defined as:

$$d_{ij} = 1 - \left(\frac{\#\text{ of common nodes between i and j}}{\frac{1}{2}(\#\text{ of nodes for language i} + \#\text{ of nodes for language j})}\right)^{\lambda}. \tag{7}$$

As no theoretical basis has been established for choosing the correct value of λ, following Fearon (2003), we fix the value of λ equal to 0.5 in our analysis.[12]

We can now calculate a weighted measure of average distance of a country's population from the official language. The data on the number and size of linguistic groups in the country comes from Fearon (2003), which takes into account all linguistic groups that form at least 1% of the population share.[13] The average distance from the official language (ADOL) for any country i is calculated as:

$$ADOL_i = \sum_{j=1}^{n} P_{ij} d_{jo}, \tag{8}$$

where n are the number of linguistic groups in the country, P_{ij} refers to the population share of group j in country i and d_{jo} refers to the distance of group j from the official language.[14] To test the claim that choosing colonial languages increases ADOL, the following reduced form regression is implemented:

$$ADOL_i = \alpha + \delta_1 Colonial\ Language_i, + \beta X_i + \varepsilon_i, \tag{9}$$

where $ADOL_i$ is the index measuring average distance from the official language for country i. $Colonial\ Language_i$ is a dummy indicating whether the country choose a colonial language not

[12] We also re-do our analysis using multiple values of λ that have been used in the literature. Our results remain qualitatively very similar and are available on request.

[13] Fearon's (2003) classification of groups, relying on a range of secondary sources, has been recognized in the literature as both principled and objective. See Esteban et al. (2012) for a discussion of the same.

[14] For details on the coding rules when there is more than one official language refer to Laitin and Ramachandran (2014).

belonging to any major indigenous group in the country and X_i is a vector of controls.[15] The results of the estimation exercise are shown in Table II.

In column (1) the dummy for having a colonial language is seen to be positive and statistically significant at the 1 percent level, indicating that choosing a colonial language increases ADOL by 0.64. In column (2), we additionally control for the index of state history from the work of Bockstette et al. (2002). Controlling for the state antiquity index does not affect either the significance or the magnitude of the coefficient.[16] Finally, column (3) includes continent dummies; given language policy choices are closely correlated to continent dummies, not surprisingly the coefficient on the colonial language dummy drops though it remains statistically significant at the 1 percent level. The results presented in Table II provides evidence for the claim that choosing colonial languages results in increasing the ADOL.

The reason why we care about distance from the official language is based upon the evidence presented in Laitin and Ramachandran (2014); here we provide a sketch of the argument and refer the interested reader to Laitin and Ramachandran (2014) for detailed exploration of the relationship between official language choice and socio-economic development.

The distance from official language is assumed to affect socio-economic outcomes through two specific channels - (i) the individual's *distance* from the official language (ii) individual's *exposure* to the official language. More concretely, it is assumed that as distance to the official language *increases* and exposure to the official language *decreases*, the learning costs associated with obtaining human capital increase in society. In addition, the use of a distant language increases the cost of acquiring and processing pertinent health information, and acts as a barrier to fostering desirable health behavior, as well in affecting access and quality of health care provided. These differences in physical and mental human capital in turn translate into differences in productivity and wealth. Thus choosing as official a language that is distant from the

[15]The coding rule followed is that if a country chooses a colonial language, which is spoken by less than 10 percent of the population as their mother tongue we code it as a one and zero otherwise.

[16]A formal test for equality of the coefficients in column (1) and (2) of Table II is not rejected at conventional significance levels ($z =$).

indigenous languages, whose use is severely restricted in day to day interactions (in other words retaining a colonial language) has negative consequences on the levels of socio-economic development.

The above argument implies that increasing linguistic diversity will make it more likely that countries will retain colonial language to act as official, which in turn based upon the evidence presented in Table II will imply an increase in ADOL. Figure III provides graphic evidence for this relationship. Panel A of Figure III shows the scatter plot and the fitted line between

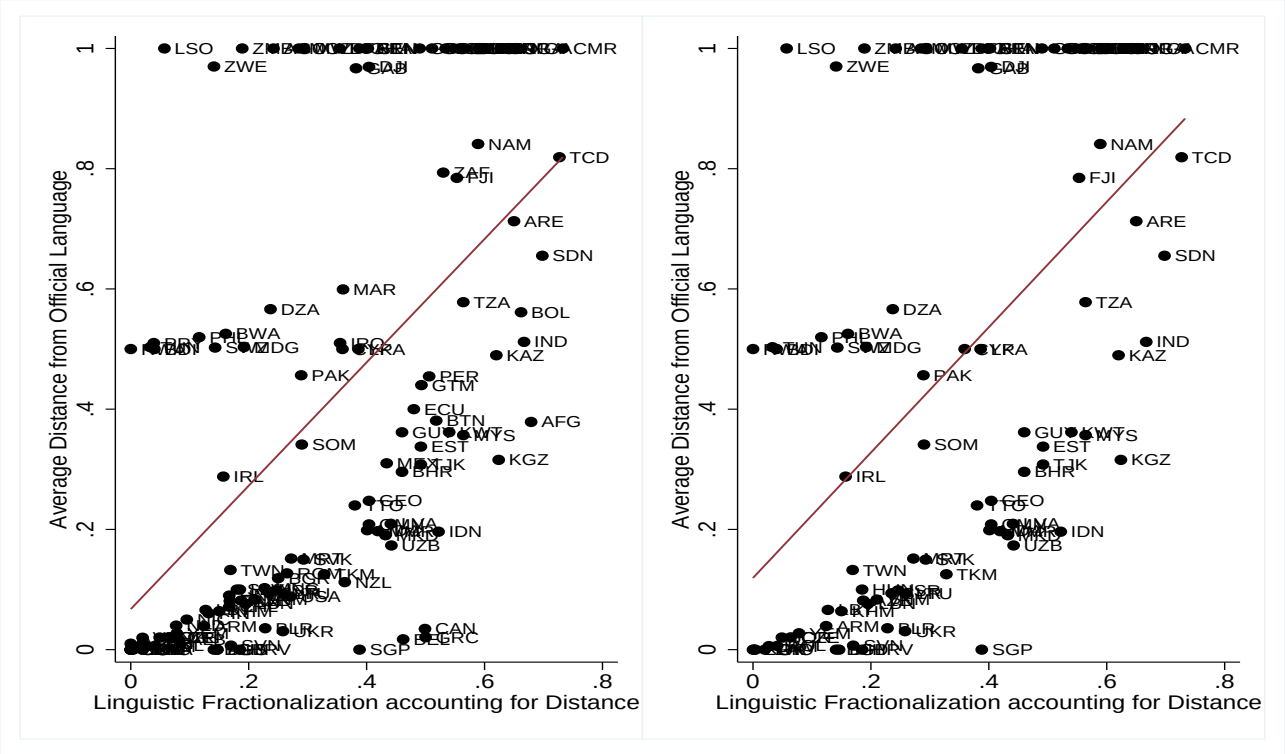

Figure I: **Ethnolinguistic fractionalization and average distance from official language**

the Greenberg index of linguistic diversity, accounting for structural distance between group's languages taken from the work of Fearon (2003), and the distance from the official language for the sample of countries that have ever been colonies.[17] Panel B in turns plots the same relationship but for reasons of comparability considers only the sample of countries that gained

[17] As our theory speaks directly to the conundrum facing post-colonial states, we consider the sample of countries that have ever been colonies. The data on whether a country was ever a colony comes from Treisman (2007)

independence post-1945. In both panels we can see that as linguistic diversity increases, consistent with the conceptual framework presented before, the average distance from the official language increases.

3.2 Ordinary least square estimates

In order to empirically test our theory we estimate an OLS regression given by:

$$ADOL_i = \alpha + \delta_1 Writing\ Tradition_i + \delta_2 Linguistic\ Fractionalization_i + \beta X_i + \varepsilon_i, \qquad (10)$$

where $ADOL_i$ is the index measuring average distance from the official language for country i. $Writing\ Tradition_i$ is a dummy indicating whether the country had a standardized writing scripting for a major indigenous group at least a generation before independence, whereas $Linguistic\ Fractionalization_i$ is an index measuring the levels of linguistic diversity and X_i is a vector of controls.

The results are shown in Table III. The OLS estimates in column (1), where we regress ADOL on the two hypothesized explanatory factors - the dummy for having a written tradition and the Greenberg index of linguistic fractionalization from the work of Fearon (2003) - provide strong support for our theoretical framework. Not only are the estimates statistically significant but also explain more than 80 percent of the variation observed in ADOL.

Insert Table III

Columns (2) and (3) additionally control for log GDP per capita at independence, and the log of population in 1500 CE, respectively. Inclusion of controls which measure wealth, either at the stroke of independence when language policy choices were instituted, or in the Middle Ages, is to explore for the relative importance of the role of wealth or stage of development compared to the factors emphasized by our theory. Both log GDP per capita and log population in 1500 CE are not only statistically insignificant, but the standardized coefficients are close to zero and

are of the wrong sign. In column (4) we show that our results are robust to the inclusion of continent dummies.

In column (5), instead of the Greenberg index we use a measure of linguistic fractionalization from the work of Alesina et al. (2003) which does not account for distance. We see that even this alternative measure is a statistically significant and economically meaningful predictor of distance from official language.[18]

3.3 An instrumental variable approach

The OLS estimates provide strong support in favor of the proposed theoretical framework in section 2. However, one concern is that the use of a dummy to capture whether a country had a writing tradition or not is beset with a host of endogeneity problems. Countries which possessed a writing tradition compared to those that did not conceivably differ on many other important characteristics. Thus the regressions presented in Table III are subject to the criticism that our dummy variable is in fact capturing these other unobservable characteristics correlated with possessing a writing tradition.

To address this concern we rely on using an instrument that is correlated with having a writing tradition, but plausibly uncorrelated with any other country characteristics that potentially affects the choice of official language. Drawing on the work of Diamond (1997), we use distance from the sites at which writing was independently invented as an instrument for possessing a writing tradition. He contends that writing was invented in societies where certain prerequisites were satisfied; first, stratified societies with complex and centralized political institutions where writing was useful for bureaucratic and administrative purposes; and second, societies with social and agricultural mechanisms for generating the food surpluses required to feed scribes. Such conditions were satisfied in three societies, where writing is said to have

[18] It should be noted that we include the Herfindahl index of ELF only for the sake of completeness, as our theory is directed towards indices of ELF which account for distance.

been originally invented, Mesopotamia (Sumer) around 3200 BCE, in China around 1200 BCE, and in Mesoamerica around 600 BCE. The rest of the world acquired the writing tradition later through trade, conquest and contact with the societies where writing was invented.[19] He argues that geography was a crucial factor as to why Tonga's maritime proto-empire, the Hawaiian state emerging in the late 18th century, all of the states and chiefdoms of subequatorial Africa and sub-Saharan West Africa, and the largest native North American societies, those of the Mississippi Valley and its tributaries, did not acquire writing before the expansion of Islam and the arrival of the Europeans.

The instrument thus exploits the exogenous component for the probability of having a writing tradition, i.e. geography. The key underlying assumption for it to be a valid instrument is that the distance from these sites of invention should have no independent impact on official language choice, except through the channel of affecting the probability of possessing a writing tradition. We operationalize the measure by calculating the Great-Circle-Distance, using the Haversine formula, from each of the sites of invention to every country in our sample. We then take the minimum of the distance from the three sites, as the measure of distance from the place of invention of writing.

Table IV shows the results of the IV regression. In Panel B are shown the results of the first stage regression of writing tradition from the minimum of the distance from the sites of invention of writing. The minimum distance from the sites of invention of writing is seen to a statistically significant predictor of possessing a writing tradition. The F-statistics of the first stage regression lie in the range of 14-44, implying that the proposed instrument is not weak.

Insert Table IV

In Panel A are shown the results of the second stage regression of ADOL on the Greenberg index of linguistic fractionalization, and writing tradition instrumented with the minimum of

[19]He discusses two primary forms in which other states acquired writing - blueprint copying and idea diffusion. Refer to Diamond (1997, chapter 12) for further details.

the distance from the sites of invention of writing. In line with the OLS results, both linguistic fractionalization and having a writing tradition are seen to be not only statistically significant predictors of ADOL but again explain almost 80 percent of the observed variation. Again measures of past wealth - log GDP per capita at independence and log population in 1500 CE - are not only unimportant predictors of ADOL in a statistical sense, but even the point estimates are close to zero.

One concern that remains is that the distance from the sites of invention of writing also affected the development of state institutions or government quality, which in turn affect the choice of official language. To determine whether such a channel is indeed relevant we regress the distance from the sites of invention of writing on three widely used measures of institutional quality or governance, namely, (i) average protection against expropriation risk from the Political Risk Services (PRS) group averaged over the years 1995-05; (ii) social infrastructure combining government anti-diversion policies and openness to international trade from the work of Hall and Jones; and (iii) constraints on the executive from Polity-IV and averaged over the years 1960-2000. The results in Table V show that the distance measure is not a significant correlate of any of the three institutional or state quality measures, and in fact the R-squared is always less than 1 percent and the F-statistic also takes a value less than one in all three regressions. The evidence in Section 3.2 and 3.3 thus provide strong support for the presented theoretical framework and show that linguistic diversity and the availability of a well-developed writing tradition are two of the key factors affecting official language choices today.

4 Linguistic diversity and official language: Implications for the cross-country literature on diversity

Since the seminal work of Easterly and Levine (1997), there has been a growing consensus that diversity - ethnic, linguistic and religious - has negative consequences for development. This

said, the mechanism through which diversity operates remains contentious (Habyarimana et al., 2009). In this section we analyze the implications of linguistic diversity for economic development in light of the demonstrated relationship between linguistic diversity and official language choice that so far has remained unaccounted for in the cross-country empirical literature.

It is not our intention here to demonstrate the role of official language choice in affecting socio-economic development, and for evidence on this we refer the reader to Laitin and Ramachandran (2014). The main aim of this exercise, it should be emphasized, is to explore how accounting for average distance to the official language changes the explanatory power of standard indices of linguistic diversity in explaining cross country differences in redistribution (Desmet et al. 2009, Alesina et al. 2001), quality and effectiveness of government (La Porta et al. 1999), and productivity and income levels (Alesina and Ferrara 2005, Easterly and Levine 1997).

The first dependent variable considered is transfers and subsidies as a share of GDP from the work of La Porta et al. (1999). This is used as a proxy for redistribution, as the literature has argued that diversity reduces government transfers and that altruistic attitudes are more prevalent within homogenous groups than across ethnically or culturally diverse groups (Desmet et al. 2009). To put this relationship of diversity and transfers to test, we use the Greenberg (1956) index of linguistic diversity and the Fearon (2003) list of groups.

Insert Table VI

The results are shown in Table VI. In column (1), in line with previous work, the Greenberg index of linguistic diversity is seen to reduce the level of transfers and subsidies and is statistically significant at the 1 percent level. Column (2) controls for ADOL; as can be seen, controlling for ADOL not only turns the standardized coefficient on the index of linguistic diversity insignificant, the point estimate switches signs and turns positive. In column (3) we include legal origin dummies based on the work of La Porta et al. (1999), and in column (4) include dummies for Asia and Africa. In all columns the index of linguistic diversity is seen to remain insignificant,

and the point estimate positive, overturning cross country results from earlier studies. The coefficient on ADOL remains negative and significant, and the magnitude is larger than of all the other explanatory variables considered.

Insert Table VII

Tables VII and VIII consider two other variables proposed by La Porta et al. (1999) - the corruption score from the Political Risk Services Group (PRS) and the infant mortality rate in 2010 as indicator of the quality of government. We see a similar pattern to the one observed in Table VI, viz. the Greenberg index of linguistic diversity, in line with the existing literature, increases the level of corruption and the infant mortality rate.[20] However, once we control for ADOL, the coefficient on linguistic diversity again changes signs and becomes insignificant. We additionally control for legal origins and an Africa and Asia dummy and the results remain very similar.

Insert Table VIII

Finally, we consider two indicators of productivity and income - log output per worker from the work of Hall and Jones (1999) and log GDP per capita in 2005. Linguistic diversity, as measured by the Greenberg index, again has a sizeable negative impact on productivity and income level, as long as we do not control for ADOL. Once we account for ADOL, as shown in Table IX and X, once again linguistic diversity not only becomes insignificant but the point estimate turns positive suggesting diversity, if anything, has beneficial effects on productivity and income except through its effect on ADOL.

Insert Table IX

The presented results connote that the existing cross country empirical literature on linguistic diversity and economic development has been inadvertently attributing observed negative effects directly to levels of diversity whereas at least a part of the effects stem through the indirect

[20]The corruption score is on an index of 0-10, where 10 implies the lowest level of corruption.

channel of the choice of a "distant" language as the official language. It is important to stress that the presented evidence is not to claim that linguistic diversity operates exclusively through the channel of official language choice, or *taste based* and *community social sanction* mechanisms that have been highlighted in the literature are not important depending on the context; rather the motivation is to highlight the fact that average distance from official language is an important channel that has been overlooked in the existing literature and needs to be accounted for in future analysis.

Insert Table X

5 Some applications of the theoretical framework to existing empirical evidence

In this section we apply the insights from our theoretical framework to some existing studies on the relation between ethnic diversity and development. We show that our theoretical framework is able to enrich our understanding of the data, and contributes to better discern the mechanisms at work.

5.1 Application to "Tribe or Nation" (Miguel, 2004)

Miguel (2004) using a colonial-era boundary placement as a natural experiment compares local ethnic diversity and public good provision in two rural areas in Western Kenya and Western Tanzania, respectively. He finds that ethnic diversity has negative consequences in the Busia district in Kenya, though it has no, or if anything a positive effect in the Meatu district of Tanzania. He attributes the difference in outcomes to a conscious program of nation building undertaken in Tanzania, which has been lacking in Kenya. Though we are broadly in concurrence with the role given to nation building by Miguel (2004) in explaining the observed differences, we believe

that the path undertaken, or more specifically the choice of languages used for nation building is crucial, and Tanzania is in fact a good demonstration of our theory.

The main result in the paper relates to primary school funding, and analyzed through our theoretical perspective brings to fore the role of language policy. In Tanzania, and the Meatu district, the official language, as well as that of primary schooling, is Swahili.[21] On the other hand in Kenya, the official language, as well as the language of schooling after grade 1, especially in linguistically diverse regions, is English.[22] In the statistics reported by Miguel (2004), there are three major ethnic groups in the Meatu district, Sukuma (85 percent) and the Tatutru and Nyiramba (15 percent together), whereas in the Busia district in Kenya also they are three major groups, namely, Luhya (67 percent), Teso (26 percent) and Luo (5 percent). The levels of ethnolinguistic fractionalization calculated using the Herfindahl concentration index in Kenya is 0.23 and in Tanzania 0.13.

However if instead of calculating the levels of ELF, we were to calculate the distance from the official language, in the case of Tanzania we would calculate the distance from Swahili for each of the three groups. This would imply that the distance for Sukuma and Nyiramba would be 0.10, whereas for the Taturu it would be one.[23] Thus average distance from the official language for the Meatu district would be in the range of 0.10-0.23.[24] Calculating the distance from the official language in Kenya would imply that all groups have a distance of 1, as the indigenous languages comes from either the Niger-Congo or the Nilo-Saharan language family, whereas the official language English is from the Indo-European language family. Thus the average distance from the official language would be 1 for the Busia district. The values on the

[21] This is not strictly true as both English and Swahili are official languages, but however here as our focus is on primary schooling where Swahili is used as the medium of instruction we use this simplifying assumption.

[22] Refer to Albaugh (2014, pg. 257) for details on language policy in primary schooling in Kenya.

[23] The language trees associated with each language and used for calculating distances is based on Ethnologue. Swahili, Sukuma and Nyiramba all belong to the Niger-Congo family and both share 8 out of a total 10 branches with Swahili. On the other hand Taturu is from a different language family, the Nilo-Saharan, and by construction the distance is equal to 1.

[24] As the paper reports only the population share of the Sukuma, the lower and upper bound are calculated assuming that the share of Taturu tends to a minimum of 0 to a potential maximum of 15 percent, respectively.

distance from official language are quite different for the two districts, and in line with our theory. Tanzania through adoption of an indigenous language has effectively reduced distance and thus has better outcomes as compared to Kenya. Econometrically speaking, if distance from the official language was to be included as an explanatory factor, then the differences reported by Miguel between Kenya and Tanzania could plausibly be explained by our indicator.

The low exposure to English in Busia has lowered human capital of the parents who went through the school system, and they are less able to judge or monitor the school authorities. Our favored interpretation is that parents who can judge quality of schooling, and learning outcomes, might be much more motivated to engage in ensuring all members of the community contribute to the provision of the public good. Two strands of evidence provide support to such a claim. First, the data of Miguel (2004) showing the average levels of education in Busia and Meatu are 7.4 and 4.1 years may be deceiving. It suggests that human capital is higher in Kenya. However, using the Demographic and Health Survey (DHS) data from the year 2011-12 for the two countries, we observe a different picture. In the Western region of Kenya, where Busia is located, only 51 percent of the male population recorded as having between 4 to 7 years of education are able to read a complete sentence. By contrast, in the in the Shinyanga region, where Meatu is located, 76 percent of the male population recorded as having between 4 to 7 years of education are able to read a complete sentence, thus suggesting that actual level of knowledge might be higher in Tanzania. The fact that children learn in Swahili, a language understood by the parents, allows them to ascertain their child's progress and in turn value the public good. Evidence supporting this interpretation also comes from recent work by Blimpo et al. (2011), who report that although most students in their data from Gambia are unable to read or write, still more than 90 percent of the parents report as being satisfied with their child's progress. Blimpo et al. (2011, pg. 17) attribute this disconnect to the inability of the parents to hold the schools accountable and participate effectively in school management due to their lack of ability to judge their child's level of knowledge. Thus the ability of the parents to gauge the

quality of the public good (in this case primary schooling) might be an important motivating factor determining the effort communities exert towards provision of the public good.

Another potential channel might be that use of a local language which assists human capital formation in turn impacts social preferences, norms and institutions (Refer to Jakiela et al. 2014 for evidence). Along the lines suggested by Jakiela et al. (2014) that human capital fosters respect for earned property rights it might similarly create values that eschew free riding, especially when the public good is recognized to be valuable. In sum, though we agree with the importance of nation building in explaining the results of Miguel (2004), we enrich the interpretation by highlighting the fact that choice of proximate languages used and widely understood by all rather than just the choice of a common language is a greatly added value to an ideology of nation building. This is not to suggest that other factors which have contributed to nation building in Tanzania are not important, but to stress that the use of a commonly understood language might be a key input into fostering human capital, creating civic spirit and cooperation within and between groups.

5.2 An application to the work on "Artificial States" (Alesina et al., 2011)

Alesina et al. (2011) construct measures of the extent to which countries' borders may be classified as artificial, and in turn the degree of artificiality of a state. To operationalize the notion of artificial states they use two indices - (i) how borders split ethnic groups into two separate adjacent countries, or more specifically, the proportion of population in a country belonging to a partitioned ethnicity; and (ii) the straightness of land borders, using a fractal measure, under the assumption that straight land borders are more likely to be artificial.

They show that higher levels of artificiality are correlated with lower levels of GDP per capita, and especially the indicator measuring the proportion of population comprising partitioned ethnic groups is a robust correlate of GDP per capita. To rationalize the evidence they put forth the hypothesis "When states represent people put together by outsiders, these peoples

may find it more difficult to reach consensus on public goods delivery and the creation of institutions that facilitate economic development, compared to states that emerged in a homegrown way" (Alesina et al., 2011, pg. 247).

One obvious consequence of partitioning ethnicities across national borders is the associated increase in ethnolinguistic diversity. Figure II shows the relation between the first measure of artificiality and the Greenberg index of linguistic fractionalization. Not surprisingly we observe that as the percentage of population comprising partitioned groups increases, so does the level of linguistic diversity.

Our theoretical model shows that as linguistic diversity increases, the average distance from

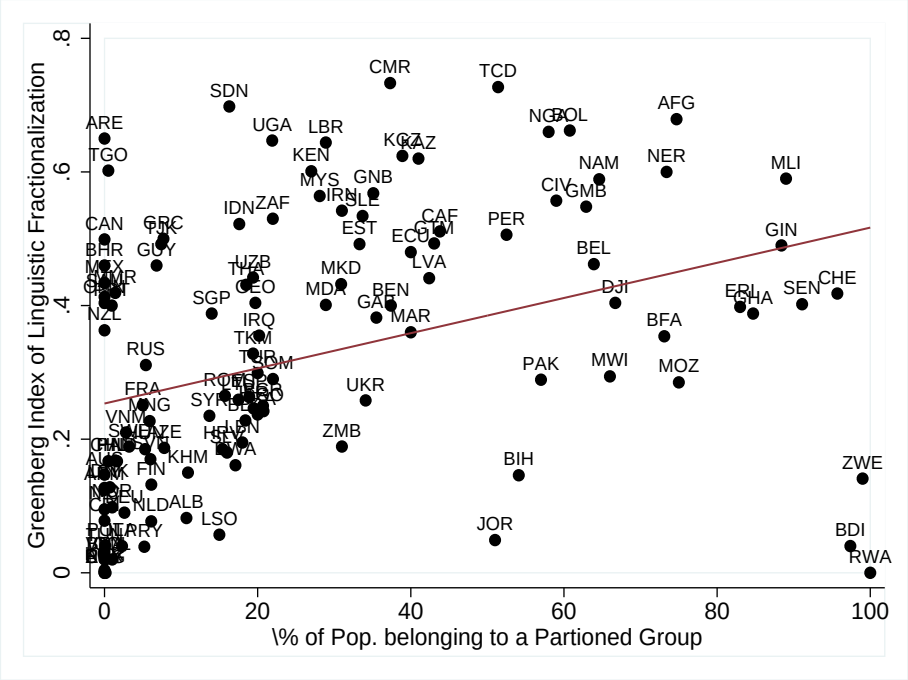

Figure II: **Measure of artificiality of states and ethnolinguistic fractionalization**

the official language increases due to the retention of the colonial language. Thus one important channel through which artificial states might be associated with poorer economic outcomes is through choosing distant languages to act as the language of education and administration, which increases the cost of human capital formation and impedes political participation and

public debate. To test our hypothesis we re-estimate Tables 6B and 6C from the work of (Alesina et al. 2011, pg. 272-73). The authors use a principal component analysis that combines three ethnic and two artificial state measures and use the first two principal components that account for the most variance as a measure of state artificiality, instead of the two measures of artificiality discussed earlier.[25]

Insert Table XI

We reproduce Tables 6B and 6C, but additionally control for the distance from the official language; the results are shown in Table XI and XII. The results show that average distance from the official language is statistically significant at the 1 percent level in 16 of the 18 regressions, whereas the second principal component is never significant and the first principal component loses significance in 5 of the 18 specifications. Moreover, comparing the standardized coefficient of the three explanatory factors shows that the magnitude of the effect of ADOL is higher than the ones predicted by the two components representing artificiality. Comparing the standardized coefficients on the first principal component, controlling and not controlling for average distance from official language, shows that controlling for ADOL reduces the magnitude on the coefficient to around half its size.[26] The results again suggest that one important channel through which creation of artificial states leads to poorer economic outcomes is through increasing linguistic diversity, which in turn increases the probability of retaining colonial languages as official. We must stress again that this is not to say that artificial states do not operate through other mechanisms to affect economic development, but to show that one important channel through which creation of artificial states has resulted in poorer development outcomes is through the channel of language policy.

Insert Table XII

[25]This is to avoid using multiple measures that capture the same underlying concept or concepts; refer to (Alesina et al., 2011, pg. 271-72) for further details.

[26]The results are not shown here and available upon request.

6 Conclusion

We presented a theoretical framework to understand the factors affecting the choice of official language in post-colonial states. The framework showed that linguistically diverse states are unable to resolve conflicts regarding which indigenous language should be chosen to act as official; and resort to maintaining the status quo of using the colonial language. The unavailability of a written indigenous language necessitates the need to invest in creating a standardized script, orthography and modern scientific vocabulary before it can be adopted to serve the role of an official language. This was modeled as imposing an additional fixed cost or creating uncertainty regarding returns and suitability of a language that has never been used before in formal domains. We show that both these factors increase the probability of a nation retaining the former colonial language to act as official. We next provided empirical evidence in favor of our theoretical framework, and showed that the highlighted factors are not only statistically significant in explaining official language choices, but explain more than 80 percent of the variation observed in the data.

The implications of this relationship between linguistic diversity and official language choice were then explored in the context of the cross-country literature on diversity and development. It is shown that once we account for official language choice in cross-country studies explaining differences in redistribution, quality of government, wealth and productivity, in contrast to earlier literature, we find that standard indices of diversity have little or no explanatory power. We contend that a large portion of the effects of linguistic diversity is mediated through the channel of official language choice; and this has been an important omitted variable in existing cross-country studies. Our interpretation suggests that use of a language that is not spoken indigenously, and is very different from the languages locally spoken, imposes high costs for human capital formation, and impedes public debate and political participation.

Finally, we applied our framework to studies on nation building and creation of strong states.

We show that a key element of nation building is not just choice of a common language, but a language that is widely understood and spoken, and facilitates human capital formation. We make much needed progress on identifying what can help create cohesive and inclusive institutions. The feature we pinpoint is the average linguistic distance for the sum of all individuals in a society between their indigenous languages and the official language of the state. We recognize that lowering this distance is heavily constrained by history, by difficulties in compensating losers, and by elite returns to the status quo. Nonetheless, this feature, unlike ethnic linguistic fractionalization, is amenable to policy choices and therefore important for future discussions of human development.

References

Acemoglu, D. and J. Robinson (2012). *Why Nations Fail: The Origins of Power, Prosperity and Poverty*. London, Profile.

Adegbija, E. E. (1994). *Language attitudes in Sub-Saharan Africa: A sociolinguistic overview*, Volume 103. Multilingual matters.

Albaugh, E. A. (2014). *State-Building and Multilingual Education in Africa*. Cambridge University Press.

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

Alesina, A., W. Easterly, and J. Matuszeski (2011). Artificial states. *Journal of the European Economic Association 9*(2), 246–277.

Alesina, A. and E. L. Ferrara (2005). Ethnic diversity and economic performance. *Journal of Economic Literature 43*(3), 762–800.

Alesina, A., E. Glaeser, and B. Sacerdote (2001). Why doesn't the US have a European-style welfare system?

Alisjahbana, S. T. (1962). *Indonesian language and literature: two essays*. Number 11. Yale University, Southeast Asia Studies.

Anderson, B. R. O. (1990). *Language and power: Exploring political cultures in Indonesia*. Cornell University Press.

Annamalai, E. (1979). Language movements against Hindi as an official language. *Language Movements in India*.

Austin, G. (1999). *The Indian constitution: Cornerstone of a nation*. Oxford University Press, USA.

Besley, T. and T. Persson (2010). State capacity, conflict, and development. *Econometrica 78*(1), 1–34.

Besley, T. and T. Persson (2011a). Fragile states and development policy. *Journal of the European Economic Association 9*(3), 371–398.

Besley, T. and T. Persson (2011b). *Pillars of prosperity: The political economics of development clusters*. Princeton University Press.

Blimpo, M. P., N. Lahire, and D. K. Evans (2011). School-based management and educational outcomes: lessons from a randomized field experiment. *Unpublished manscript*.

Bockstette, V., A. Chanda, and L. Putterman (2002). States and markets: The advantage of an early start. *Journal of Economic Growth 7*(4), 347–369.

Bourdieu, P. (1991). *Language and symbolic power*. Harvard University Press.

Chandra, B. (2000). *India after independence: 1947-2000*. Penguin UK.

Cole, H. L., G. J. Mailath, and A. Postlewaite (1995). Incorporating concern for relative wealth into economic models. *Federal Reserve Bank of Minneapolis Quarterly Review 19*(3), 12–21.

Dardjowidjojo, S. (1998). Strategies for a successful national language policy: The Indonesian case. *International Journal of the Sociology of Language 130*(1), 35–48.

De Swaan, A. (2013). *Words of the world: The global language system*. John Wiley & Sons.

Desmet, K., S. Weber, and I. Ortuño-Ortín (2009). Linguistic diversity and redistribution. *Journal of the European Economic Association 7*(6), 1291–1318.

Diamond, J. M. (1997). *Guns, germs and steel: a short history of everybody for the last 13,000 years*. Random House.

Easterly, W. and R. Levine (1997). Africa's growth tragedy: policies and ethnic divisions. *The Quarterly Journal of Economics*, 1203–1250.

Errington, J. J. (1998). *Shifting languages*. Number 19. Cambridge University Press.

Esteban, J., L. Mayoral, and D. Ray (2012). Ethnicity and conflict: An empirical study. *The American Economic Review 102*(4), 1310–1342.

European Commission, Directorate General for Translation (2011). Lingua franca: Chimera or reality? Technical report.

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

Greenberg, J. H. (1956). The measurement of linguistic diversity. *Language*, 109–115.

Habyarimana, J., M. Humphreys, D. N. Posner, and J. M. Weinstein (2007). Why does ethnic diversity undermine public goods provision? *American Political Science Review 101*(04), 709–725.

Habyarimana, J., M. Humphreys, D. N. Posner, and J. M. Weinstein (2009). *Coethnicity: diversity and the dilemmas of collective action*. Russell Sage Foundation.

Hall, R. E. and C. I. Jones (1999). Why do some countries produce so much more output per worker than others? *The Quarterly Journal of Economics 114*(1), 83–116.

Harsanyi, J. C. and R. Selten (1988). A general theory of equilibrium selection in games. *MIT Press Books 1*.

Heffetz, O. and R. H. Frank (2008). Preferences for status: Evidence and economic implications. *Handbook of Social Economics, Jess Benhabib, Alberto Bisin, Matthew Jackson, eds 1*, 69–91.

Irschick, E. F. (1986). *Tamil Revivalism in the 1930s*. Cre-A.

Jakiela, P., E. Miguel, and V. L. te Velde (2014). You've earned it: estimating the impact of human capital on social preferences. *Experimental Economics*, 1–23.

La Porta, R., F. Lopez-de Silanes, A. Shleifer, and R. Vishny (1999). The quality of government. *Journal of Law, Economics, and organization 15*(1), 222–279.

Laitin, D. D. (1994). The tower of babel as a coordination game: Political linguistics in Ghana. *American Political Science Review 88*(03), 622–634.

Laitin, D. D. (2000). Language conflict and violence: the straw that strengthens the camel's back. *European Journal of Sociology 41*(01), 97–137.

Laitin, D. D. (2004). Language policy and civil war. *Cultural diversity versus economic solidarity. Brussels: De Boeck Université*, 171–88.

Laitin, D. D. and R. Ramachandran (2014). Language policy and human development.

Lewis, P., G. Simon, and C. Fennig (2014). *Ethnologue: Languages of the World, Seventeenth edition. Dallas, Texas: SIL International*.

Miguel, E. (2004). Tribe or nation? Nation building and public goods in Kenya versus Tanzania. *World Politics 56*(03), 328–362.

More, J. P. (1997). *The political evolution of Muslims in Tamilnadu and Madras, 1930-1947*. Orient Blackswan.

Ndamba, G. T. (2008). Mother tongue usage in learning: An examination of language preferences in Zimbabwe. *The Journal of Pan African Studies 2*(4), 171–188.

Paauw, S. (2009). One land, one nation, one language: An analysis of Indonesia's national language policy. *University of Rochester Working Papers in the Language Sciences 5*(1), 2–16.

Ramachandran, S. (1975). Anna speaks: At the Rajya Sabha, 1962-66.

Ramaswamy, S. (1997). *Passions of the tongue: language devotion in Tamil India, 1891-1970*, Volume 29. Univ of California Press.

Treisman, D. (2007). What have we learned about the causes of corruption from ten years of cross-national empirical research? *Annual Review of Political Science 10*, 211–244.

Vickers, A. (2013). *A history of modern Indonesia*. Cambridge University Press.

Weber, M. (1978). *Economy and society*. Berkeley: University of California Press.

Table II: **Regressions of colonial language dummy on average distance from the official language**

	(1)	(2)	(3)
Dummy for Colonial Language being Official	0.647***	0.631***	0.382***
	(0.0434)	(0.0488)	(0.101)
	[0.823]	[0.798]	[0.483]
State Antiquity Index		-0.176*	-0.209*
		(0.0907)	(0.116)
		[-0.0990]	[-0.117]
Continent Dummies	No	No	Yes
Observations	132	118	118
R-squared	0.677	0.695	0.736

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table III: **Regressions of writing tradition and linguistic fractionalization on average distance from the official language**

	(1)	(2)	(3)	(4)	(5)
Dummy for whether country has a written tradition	-0.612***	-0.598***	-0.613***	-0.413***	-0.358***
	(0.0375)	(0.0408)	(0.0380)	(0.0740)	(0.0635)
	[-0.728]	[-0.711]	[-0.728]	[-0.490]	[-0.424]
Linguistic fractionalization accounting for distance	0.655***	0.667***	0.646***	0.615***	
	(0.0752)	(0.0768)	(0.0779)	(0.0750)	
	[0.366]	[0.373]	[0.360]	[0.343]	
Log GDP per capita at independence in 1990 US		-0.0186		0.0205	0.0389*
		(0.0148)		(0.0201)	(0.0222)
		[-0.0453]		[0.0500]	[0.0955]
Log Population in 1500 CE			0.00738	0.00834	0.00762
			(0.00793)	(0.00962)	(0.00953)
			[0.0340]	[0.0384]	[0.0349]
Linguistic fractionalization n/actg. for distance					0.513***
					(0.0670)
					[0.389]
Continent Dummies	No	No	No	Yes	Yes
Observations	131	131	130	130	126
R-squared	0.815	0.817	0.816	0.846	0.848

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table IV: **IV Regressions of writing tradition and linguistic fractionalization on average distance from the official language**

	(1)	(2)	(3)	(4)
Panel A: Two-Stage Least Squares - Dependent variable ADOL				
Dummy for whether country has a written tradition	-0.74***	-0.75***	-0.75***	-0.82**
	(0.081)	(0.10)	(0.083)	(0.41)
	[-0.88]	[-0.89]	[-0.89]	[-0.97]
Linguistic fractionalization accounting for distance	0.58***	0.57***	0.56***	0.54***
	(0.087)	(0.097)	(0.089)	(0.11)
	[0.32]	[0.32]	[0.31]	[0.30]
Log GDP per capita at independence in 1990 US		0.0068		0.034
		(0.024)		(0.027)
		[0.016]		[0.083]
Log Population in 1500 CE			0.0079	0.021
			(0.0089)	(0.016)
			[0.036]	[0.096]
Continent Dummies	No	No	No	Yes
Observations	131	131	130	130
R-squared	0.795	0.792	0.793	0.785
Panel B: First-Stage for Writing Tradition				
Distance from Site of Invention of Writing	-0.000090***	-0.000075***	-0.000089***	-0.000027*
	(0.000017)	(0.000017)	(0.000017)	(0.000014)
	[-0.42]	[-0.35]	[-0.42]	[-0.12]
Linguistic fractionalization accounting for distance	-0.49***	-0.53***	-0.50***	-0.17
	(0.16)	(0.16)	(0.17)	(0.11)
	[-0.23]	[-0.25]	[-0.23]	[-0.079]
Log GDP per capita at independence in 1990 US		0.12***		0.024
		(0.037)		(0.031)
		[0.25]		[0.049]
Log Population in 1500 CE			-0.0055	0.028**
			(0.020)	(0.014)
			[-0.021]	[0.11]
Continent Dummies	No	No	No	Yes
Observations	131	131	130	130
R-squared	0.251	0.311	0.253	0.745
F-Stat	21.5	19.1	14.2	44.2

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table V: **IV Falsification test - Regressions of distance from sites of invention of writing on three measures of state institutions and governance quality**

	(1)	(2)	(3)
	Average Protection against Expropriation Risk	Social Infrastructure	Constraints on the Executive
Distance from Site of Invention of Writing	-1.8e-07	-6.0e-06	0.000075
	(7.3e-06)	(0.000010)	(0.000080)
	[-0.0024]	[-0.060]	[0.083]
Observations	110	95	130
R-squared	0.000	0.004	0.007
F-Stat	0.00060	0.34	0.88

In column (1), (2) and (3) the dependent variables are average protection expropriation risk, social infrastructure and constraints on the executive, respectively. *p < .10; **p < .05; ***p < .01. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table VI: **Regressions of linguistic fractionalization and average distance from the official language on Transfers & Subsidies as share of GDP (74-94)**

	(1)	(2)	(3)	(4)
Linguistic fractionalization accounting for distance	-8.126***	1.909	0.655	2.487
	(2.902)	(4.222)	(4.276)	(4.585)
	[-0.264]	[0.0621]	[0.0213]	[0.0809]
Average distance from official language		-9.821***	-8.677***	-11.11***
		(2.649)	(2.666)	(4.073)
		[-0.536]	[-0.474]	[-0.607]
Legal Origin - French			-3.028**	-3.577**
			(1.437)	(1.394)
			[-0.217]	[-0.256]
Legal Origin - Socialist			9.873***	8.951***
			(2.967)	(3.046)
			[0.335]	[0.304]
Legal Origin - German			-9.414***	-6.699***
			(1.700)	(2.473)
			[-0.163]	[-0.116]
Legal Origin - Scandinavian			4.873***	3.982**
			(1.538)	(1.616)
			[0.0845]	[0.0691]
Observations	68	68	68	68
R-squared	0.070	0.251	0.480	0.508

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table VII: **Regressions of linguistic fractionalization and average distance from the official language on Corruption Score from ICRG**

	(1)	(2)	(3)	(4)
Linguistic fractionalization accounting for distance	-1.773*	0.317	-0.0695	0.634
	(0.936)	(1.248)	(1.244)	(1.329)
	[-0.185]	[0.0331]	[-0.00725]	[0.0660]
Average distance from official language		-1.782**	-1.552**	-2.620**
		(0.687)	(0.723)	(1.095)
		[-0.336]	[-0.293]	[-0.495]
Legal Origin - French			-0.957*	-1.028**
			(0.494)	(0.483)
			[-0.232]	[-0.249]
Legal Origin - Socialist			0.000319	-0.0158
			(0.665)	(0.663)
			[4.73e-05]	[-0.00234]
Legal Origin - German			-1.126*	-0.555
			(0.644)	(0.728)
			[-0.0555]	[-0.0273]
Legal Origin - Scandinavian			3.679***	3.441***
			(0.575)	(0.610)
			[0.181]	[0.170]
Africa				0.614
				(0.743)
				[0.143]
Asia				-0.785
				(0.546)
				[-0.160]
Observations	96	96	96	96
R-squared	0.034	0.100	0.195	0.231

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table VIII: **Regressions of linguistic fractionalization and average distance from the official language on Infant Mortality Rate in 2010**

	(1)	(2)	(3)	(4)
Linguistic fractionalization accounting for distance	71.19***	-24.95	-22.14	-3.112
	(18.89)	(16.91)	(16.90)	(17.25)
	[0.328]	[-0.115]	[-0.102]	[-0.0143]
Average distance from official language		93.97***	92.59***	46.01***
		(8.070)	(9.074)	(17.22)
		[0.773]	[0.761]	[0.378]
Legal Origin - French			9.419	8.728
			(7.482)	(6.812)
			[0.103]	[0.0953]
Legal Origin - Socialist			4.269	5.744
			(7.539)	(6.373)
			[0.0402]	[0.0541]
Legal Origin - German			-8.480	-16.14*
			(8.610)	(8.507)
			[-0.0162]	[-0.0308]
Legal Origin - Scandinavian			-12.26	-4.513
			(7.403)	(6.754)
			[-0.0234]	[-0.00860]
Africa				45.67***
				(13.13)
				[0.475]
Asia				15.02***
				(5.382)
				[0.146]
Observations	131	131	131	131
	0.518	0.582		

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table IX: **Regressions of linguistic fractionalization and average distance from the official language on Log Output per Worker**

	(1)	(2)	(3)	(4)
Linguistic fractionalization accounting for distance	-1.545***	0.634	0.584	0.287
	(0.391)	(0.410)	(0.394)	(0.365)
	[-0.332]	[0.136]	[0.126]	[0.0620]
Average distance from official language		-2.039***	-2.099***	-1.185***
		(0.215)	(0.202)	(0.335)
		[-0.789]	[-0.815]	[-0.460]
Legal Origin - French			-0.224	-0.248
			(0.161)	(0.153)
			[-0.108]	[-0.120]
Legal Origin - Socialist			-0.873**	-0.986***
			(0.415)	(0.369)
			[-0.193]	[-0.218]
Legal Origin - German			0.0602	0.292
			(0.194)	(0.218)
			[0.00608]	[0.0295]
Legal Origin - Scandinavian			0.771***	0.587***
			(0.175)	(0.189)
			[0.0778]	[0.0592]
Africa				-0.913***
				(0.271)
				[-0.445]
Asia				-0.399*
				(0.224)
				[-0.143]
Observations	94	94	93	93
R-squared	0.110	0.514	0.556	0.615

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table X: **Regressions of linguistic fractionalization and average distance from the official language on Log GDP per capita in 2005**

	(1)	(2)	(3)	(4)
Linguistic fractionalization accounting for distance	-1.362***	0.993*	0.966*	0.830*
	(0.501)	(0.563)	(0.564)	(0.495)
	[-0.233]	[0.170]	[0.165]	[0.142]
Average distance from official language		-2.254***	-2.427***	-1.630***
		(0.266)	(0.296)	(0.464)
		[-0.691]	[-0.744]	[-0.499]
Legal Origin - French			-0.521**	-0.531**
			(0.220)	(0.212)
			[-0.212]	[-0.216]
Legal Origin - Socialist			-0.733***	-0.746***
			(0.277)	(0.248)
			[-0.252]	[-0.256]
Legal Origin - German			0.542**	1.033***
			(0.260)	(0.297)
			[0.0393]	[0.0748]
Legal Origin - Scandinavian			0.864***	0.573**
			(0.233)	(0.221)
			[0.0626]	[0.0415]
Africa				-0.970***
				(0.342)
				[-0.375]
Asia				-0.752***
				(0.216)
				[-0.270]
Observations	126	126	126	126
R-squared	0.054	0.369	0.430	0.502

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table XI: **Impact on Log GDP per Capita of First Two Principal Components Controlling for Continent Dummies and Average Distance from Official Language.**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Average distance from official language	-1.30***	-1.43***	-0.97*	-0.81	-1.75***	-1.41***	-1.71***	-1.46***	-1.14***
	(0.39)	(0.38)	(0.50)	(0.58)	(0.45)	(0.39)	(0.34)	(0.41)	(0.31)
	[-0.46]	[-0.50]	[-0.34]	[-0.29]	[-0.62]	[-0.50]	[-0.60]	[-0.51]	[-0.40]
First principal component	0.32***	0.23**	0.066	0.25**	0.18	0.23**	-0.059	0.22*	0.32***
	(0.10)	(0.11)	(0.13)	(0.11)	(0.12)	(0.11)	(0.15)	(0.12)	(0.086)
	[0.40]	[0.28]	[0.082]	[0.31]	[0.22]	[0.29]	[-0.073]	[0.27]	[0.40]
Second principal component	0.019	0.038	-0.080	0.023	0.0039	0.037	-0.075	0.040	0.052
	(0.089)	(0.078)	(0.077)	(0.076)	(0.083)	(0.078)	(0.078)	(0.078)	(0.079)
	[0.019]	[0.037]	[-0.078]	[0.022]	[0.0038]	[0.036]	[-0.073]	[0.039]	[0.051]
Climate, zone A (hot, rainy)		-0.50*	-0.28	-0.54**	-0.39	-0.52*	-0.36	-0.52*	-0.39
		(0.26)	(0.28)	(0.27)	(0.27)	(0.27)	(0.24)	(0.28)	(0.26)
		[-0.17]	[-0.093]	[-0.18]	[-0.13]	[-0.17]	[-0.12]	[-0.17]	[-0.13]
Africa			-1.73***	-0.51					
			(0.46)	(0.34)					
			[-0.74]	[-0.22]					
Latin America			-1.44***		-0.39				
			(0.31)		(0.25)				
			[-0.53]		[-0.14]				
Asia and Oceania			-1.08**			0.11			
			(0.48)			(0.32)			
			[-0.24]			[0.025]			
Europe			-0.31				1.14***		
			(0.50)				(0.37)		
			[-0.11]				[0.39]		
Middle East			-1.15**					-0.10	
			(0.45)					(0.41)	
			[-0.20]					[-0.018]	
North America									1.67***
									(0.28)
									[0.17]
Observations	71	71	71	71	71	71	71	71	71
R-squared	0.668	0.690	0.763	0.702	0.703	0.691	0.735	0.690	0.714

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table XII: **Impact on Log GDP per Capita of First Two Principal Components Controlling for Other Development Determinants and Average Distance from Offical Language.**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Average distance from official language	-1.28***	-1.30***	-1.36***	-1.29***	-1.35***	-1.45***	-1.42***	-1.24***	-1.37***
	(0.41)	(0.33)	(0.27)	(0.40)	(0.37)	(0.37)	(0.39)	(0.31)	(0.44)
	[-0.45]	[-0.46]	[-0.48]	[-0.46]	[-0.48]	[-0.51]	[-0.50]	[-0.44]	[-0.48]
First principal component	0.32***	0.15	0.0082	0.32***	0.28**	0.24*	0.23**	0.29***	0.24**
	(0.11)	(0.11)	(0.088)	(0.11)	(0.11)	(0.12)	(0.11)	(0.083)	(0.11)
	[0.39]	[0.18]	[0.010]	[0.39]	[0.34]	[0.29]	[0.29]	[0.35]	[0.29]
Second principal component	0.015	0.042	0.015	-0.00090	0.076	0.038	0.040	0.056	0.040
	(0.088)	(0.073)	(0.058)	(0.094)	(0.077)	(0.078)	(0.082)	(0.077)	(0.077)
	[0.014]	[0.041]	[0.015]	[-0.00088]	[0.074]	[0.037]	[0.039]	[0.055]	[0.039]
Climate, zone B (hot, dry)	-0.13								
	(0.30)								
	[-0.033]								
Tropics (%)		-0.76***							
		(0.23)							
		[-0.32]							
Distance to Equator			2.96***						
			(0.59)						
			[0.48]						
Desert (%)				-0.34					
				(0.28)					
				[-0.075]					
Climate, zone A (hot, rainy)					-0.43*	-0.51*	-0.50*	-0.51*	-0.51*
					(0.25)	(0.26)	(0.26)	(0.26)	(0.27)
					[-0.15]	[-0.17]	[-0.17]	[-0.17]	[-0.17]
Land Area					0.094				
					(0.061)				
					[0.11]				
Population Density						-0.022			
						(0.084)			
						[-0.025]			
Trade propensity							-0.011		
							(0.15)		
							[-0.0059]		
English-speaking share								1.27	
								(0.79)	
								[0.13]	
European language-speaking share									0.070
									(0.22)
									[0.024]
Observations	71	71	71	70	71	71	71	71	71
R-squared	0.669	0.726	0.766	0.666	0.700	0.691	0.690	0.705	0.691

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Appendix

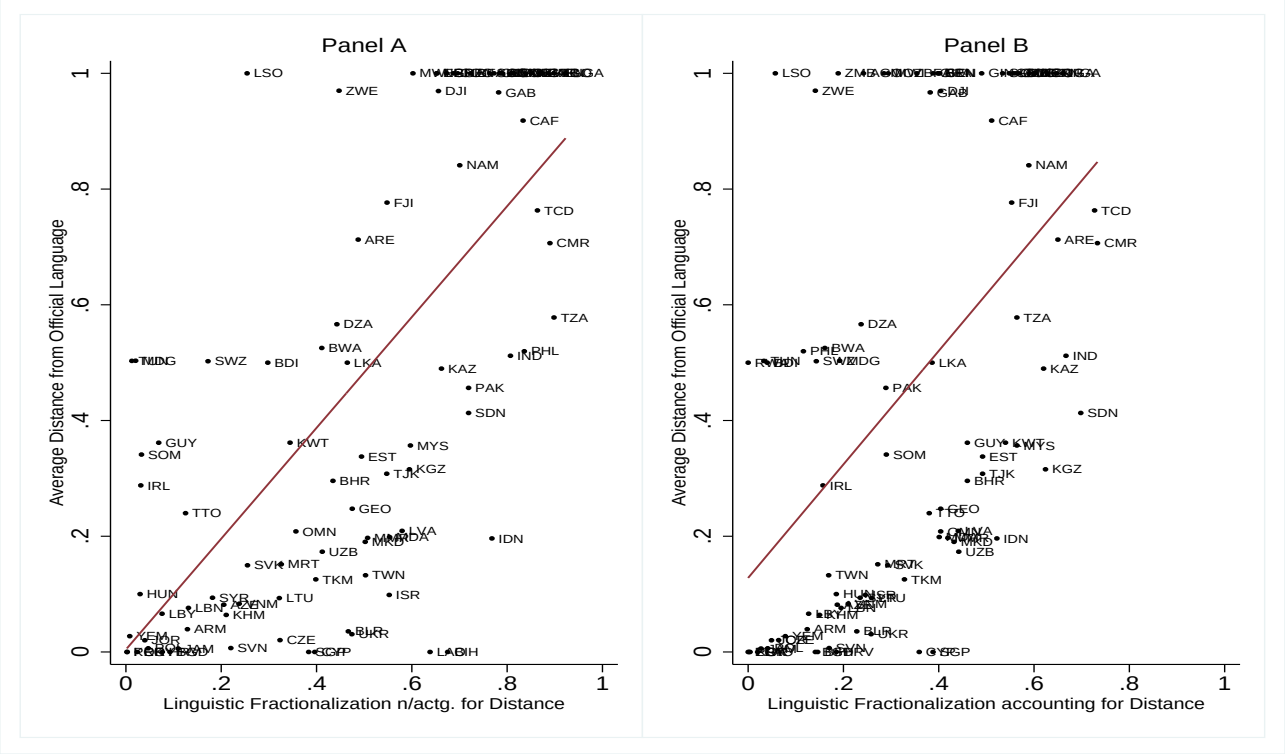

Figure III: **Ethnolinguistic fractionalization and ADOL - Sample of countries obtaining independence post-1945**

Table XIII: **Regressions of writing tradition and linguistic fractionalization on ADOL - Sample of countries obtaining independence post-1945**

	(1)	(2)	(3)	(4)	(5)
Dummy for whether country has a written tradition	-0.610***	-0.582***	-0.610***	-0.468***	-0.398***
	(0.0397)	(0.0476)	(0.0400)	(0.0815)	(0.0719)
	[-0.760]	[-0.725]	[-0.760]	[-0.583]	[-0.492]
Linguistic fractionalization accounting for distance	0.646***	0.664***	0.646***	0.605***	
	(0.0926)	(0.0923)	(0.0939)	(0.0868)	
	[0.340]	[0.349]	[0.340]	[0.318]	
Log GDP per capita at independence in 1990 US		-0.0252		0.0112	0.0559*
		(0.0188)		(0.0305)	(0.0299)
		[-0.0649]		[0.0288]	[0.144]
Log Population in 1500 CE			0.00648	0.00899	0.00814
			(0.00898)	(0.0117)	(0.0107)
			[0.0311]	[0.0432]	[0.0391]
Linguistic fractionalization n/actg. for distance					0.514***
					(0.0710)
					[0.377]
Observations	94	94	94	94	93
R-squared	0.832	0.835	0.833	0.861	0.884

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.

Table XIV: **IV Regressions of writing tradition and linguistic fractionalization on ADOL - Sample of countries obtaining independence post-1945**

	(1)	(2)	(3)	(4)	(5)
Panel A: Two-Stage Least Squares - Dependent variable ADOL					
Dummy for whether country has a written tradition	-0.71***	-0.72***	-0.71***	-0.65**	-0.56**
	(0.062)	(0.092)	(0.062)	(0.25)	(0.26)
	[-0.88]	[-0.90]	[-0.89]	[-0.81]	[-0.69]
Linguistic fractionalization accounting for distance	0.58***	0.57***	0.58***	0.56***	
	(0.094)	(0.10)	(0.094)	(0.11)	
	[0.31]	[0.30]	[0.31]	[0.29]	
Log GDP per capita at independence in 1990 US		0.0088		0.021	0.061**
		(0.028)		(0.030)	(0.025)
		[0.023]		[0.053]	[0.16]
Log Population in 1500 CE			0.0063	0.015	0.014
			(0.0094)	(0.014)	(0.013)
			[0.030]	[0.073]	[0.066]
Linguistic fractionalization n/actg. for distance					0.46***
					(0.10)
					[0.34]
Observations	94	94	94	94	93
R-squared	0.818	0.814	0.818	0.848	0.875
Panel B: First-Stage for Writing Tradition					
Distance from Site of Invention of Writing	-0.00015***	-0.00012***	-0.00015***	-0.000054***	-0.000048**
	(0.000021)	(0.000023)	(0.000022)	(0.000020)	(0.000020)
	[-0.60]	[-0.46]	[-0.60]	[-0.21]	[-0.19]
Linguistic fractionalization accounting for distance	-0.28	-0.36*	-0.28	-0.18	
	(0.20)	(0.19)	(0.20)	(0.14)	
	[-0.12]	[-0.15]	[-0.12]	[-0.077]	
Log GDP per capita at independence in 1990 US		0.15***		0.012	0.00033
		(0.041)		(0.043)	(0.040)
		[0.30]		[0.026]	[0.00069]
Log Population in 1500 CE			-0.0080	0.026	0.028*
			(0.021)	(0.016)	(0.016)
			[-0.031]	[0.099]	[0.11]
Linguistic fractionalization n/actg. for distance					-0.27***
					(0.10)
					[-0.16]
Observations	94	94	94	94	93
R-squared	0.407	0.480	0.408	0.775	0.787
F-Stat	31.3	27.7	20.7	36.6	38.7

$*p < .10; **p < .05; ***p < .01$. Robust SE's in parenthesis and standardized coefficients in square brackets.