Bilingualism, Language Shift, and Industrialization in Mid-20$^{\text{th}}$ Century India[*]

David Clingingsmith[†]
Department of Economics
Case Western Reserve University

April 4, 2008

Abstract

Economists have linked linguistic heterogeneity to poor economic outcomes, including low economic growth. Linguists have noted that linguistic heterogeneity has been declining worldwide for centuries. This paper argues that an important factor behind the consolidation of languages has been the rise of economic activities in which communication is relatively important, such as manufacturing and services. When it is valuable to be able to communicate widely, there is an incentive to become bilingual, particularly for speakers of minority languages. The children of bilinguals may then assimilate to the bigger language, producing consolidation. I use a simple framework to illustrate relationships between factory employment, bilingualism, and language consolidation. I then explore these relationships empirically using a panel of Indian districts for 1931 and 1961. My instrumental variables estimates show growth of manufacturing employment strongly encouraged bilingualism in mid-20$^{\text{th}}$ century India, particularly among minority language speakers. Bilingualism among speakers of a language is correlated with its relative decline: A one standard deviation increase in manufacturing employment decreased linguistic heterogeneity by a third of a standard deviation in Indian districts.

[*] Please do not cite without permission.

[†] Email: david.clingingsmith@case.edu. I am grateful for discussions with Leah Platt Boustan, Carola Frydman, Claudia Goldin, Eric Hilt, Lakshmi Iyer, Asim Khwaja, Michael Kremer, Bob Margo, Simona Mkrtschjan, Rohini Pande, and Jeffrey Williamson. Thanks to participants at the 45$^{\text{th}}$ Cliometrics Conference, the 36$^{\text{th}}$ Annual Conference on South Asia, and seminar audiences at Harvard, UC Davis, University of Toronto, Vanderbilt, Reed College, Case Western, and University of British Columbia for useful comments. A portion of this project was conducted while I was a fellow of the Project on Justice, Welfare, and Economics at the Weatherhead Center for International Affairs at Harvard University. The Harvard Department of Economics's Warburg Funds supported the data entry.

1 Introduction

Language is the primary means individuals use to communicate with one another. Speakers of a language form a network within which the cost of communication is relatively low. Lower transaction costs within languages mean that there can be an economic advantage to speaking a widely-known language. The consolidation of languages has accordingly been associated with the expansion of markets and the growth of state power (e.g. Weber 1976). Linguists have noted that widely-known languages have been growing at the expense of smaller languages worldwide since at least 1500 (Hill 1978; Krauss 1992; Crystal 1997; Gordon 2005). Many thousands of languages have gone extinct, and just 25 have become the mother tongues to 70% of the world's population.

This paper asks whether a shift in the structure of a developing economy toward communication-intensive sectors, such as manufacturing, is an important cause of language consolidation. I argue that as an economy makes the transition from agriculture to manufacturing and services, communication becomes more important to production. The scale and complexity of manufacturing in particular requires greater coordination between workers. I address this question by looking at the relationship between manufacturing and language consolidation in mid-20th century India. The paper attempts to both broaden our understanding of the economics of language and to show how the languages spoken by a population are endogenous to the process of economic development.

Language ability is an important form of human capital, particularly in developing economies. A growing literature argues that commonality of language is associated with better economic performance. Economic growth is negatively correlated with linguistic heterogeneity across countries (Alesina & La Ferrara 2005). Gravity models of trade show that common language has large positive effects on the trade flow between countries (Anderson & van Wincoop 2004). Ethnic heterogeneity in U.S. counties is negatively correlated with the provision of public goods (Alesina *et al.* 1999). This literature has generally taken language to be an exogenous attribute of a country or region over the medium term. Given that language ability can be acquired by choice and given that parents can choose which of their known languages to pass onto their children, it certainly seems likely that the distribution of language ability in a population will respond to economic incentives. Labor economists studying immigrants have argued that ability in the receiving country's language earns returns in the labor market (Berman *et al.* 2003; Bleakley & Chin 2004; Chiswick & Miller 1995; Dustmann & van Soest 2001). Language is thus likely to be endogenous to the process of economic development. My analysis spans a thirty year period, and speaks to the question of whether language can reasonably be considered exogenous over the medium term.

Social scientists have used the term *language shift* to refer to the intergenerational shift in mother tongues in favor of more widely-spoken languages that underlies language consolidation (Fishman 1964; Gal 1978). Bilingualism is a necessary condition for language shift because parents and children always share a language in common. If the mother tongue of a given lineage is to change across a generation, the parents must be bilingual and must pass on their second language to their children as a mother tongue. Language shift is thus a slow, generational process.

We can see the progress of language shift and consolidation in the long-run historical experience of India. Figure 1 shows that throughout the 20th century linguistic heterogeneity increased slowly but steadily, as did the share of the population speaking the five largest languages. The important role of bilingualism in linguistic consolidation is illustrated in Table 1. The table shows correlations between the growth rate of the number of speakers of a language between 1961 and 1991 and the initial number and number who were bilingual in 1961. At both the state and all-India levels, a language with more initial bilinguals grows more slowly controlling for the initial number of speakers. Languages with many speakers grow faster than small ones.

India is an excellent setting in which to study the relationship between industrialization and language shift. It is one of the world's most linguistically diverse countries. More than 180 distinct languages are spoken in India. Even within small geographic regions, a wide variety of mother tongues are typically spoken. Between 1931 and 1961, India had substantial increases in bilingualism, particularly among linguistic minorities, and a major shifts in the structure of employment toward manufacturing. Further, factors which confound the linkage between manufacturing employment and bilingualism are less salient in the 1931–1961 period than they will be later. While literacy expanded substantially, the completion of primary school was uncommon. The flow of migrants between regions with different majority languages was very slow. Finally, the redrawing of state boundaries along linguistic lines begins only at the very end of the period.

My empirical analysis looks at the role played by India's growing manufacturing employment in providing incentives for individuals to become bilingual and in leading to a decline in linguistic diversity. I develop a simple model linking manufacturing jobs to bilingualism and explore its predictions using a new panel dataset of Indian districts for 1931 and 1961.

My model is founded on two observations. First, the communication value of being bilingual will depend on how well people are able to sort according to their mother tongue. If people cannot sort perfectly according to language for the purposes of trade, bilingualism will be valuable. Second, the value of communication itself depends on the structure of the economy. The importance of workers being able to communicate with each other varies with the sector in which they are working and the level of technology. Factory workers labor in larger firms, perform more specialized tasks, and need to coordinate more than agricultural laborers. Factories typically find workers through the labor market rather than the family, which makes sorting by mother tongue more difficult. A worker in the market for a factory job will find employment more easily if he or she can speak the language of the median worker.

The model produces four main predictions. (1) A language that has a small population share in a district will have more bilinguals. This result rests on the assumption that there is imperfect sorting by language in the labor market. (2) Bilingualism will be greater when the relative return to manufacturing employment is greater. Individuals are more likely to be employed in factories if they can communicate with one another; given imperfect sorting, being able to communicate widely increases the chance of getting a factory job. When the relative return to factory work is high, it is more worthwhile to expand one's communication potential by becoming bilingual. (3) The incentive to become bilingual resulting from the chance to get a high-paying factory job will be larger for individuals whose mother tongue is a minority language. Bilingualism expands the communication potential of someone whose mother tongue is rare more than someone whose mother tongue is the majority language. (4) Bilingualism among mother tongue speakers of a small language may encourage its relative decline.

My district-level empirical analysis allows for language-by-district fixed effects. I develop an instrumental variable for manufacturing employment growth to mitigate simultaneity and omitted variables biases. I collect 1931 employment data for nine manufacturing subindustries, such as textiles, chemicals, and food processing. My instrumental variable is a prediction of how fast manufacturing employment in each district would have grown if each of its subindustries had grown at the average rate for the rest of the country. The instrument isolates the component of district-level manufacturing employment growth that resulted from national-level variation in employment demand by subindustry. This variation reflects subindustry final demand, relative productivity growth, tariffs, and world prices. The change in a district's manufacturing share that comes from the national fortunes of its different manufacturing subindustries is unlikely to be correlated with the district-level unobserved determinants of bilingualism.

I find that manufacturing employment growth has strong effects on bilingualism consistent with the predictions of my model. My instrumental variables regression shows that a one-point increase in the manufacturing share of employment leads to a 1.3 point increase in the bilingual share for minority language speakers and a 0.4 point increase in the bilingual share for majority language speakers. Both effects are statistically significant. They include the spill-over effects a manufacturing job has on other industries and activities. In absolute terms, these estimates imply that each new manufacturing job induces about 1.2 people to become bilingual. Manufacturing employment growth accounts for about a third of the mean change in bilingualism among minority language speakers.

Bilingualism is associated with the relative decline in the population speaking a language, at both the regional and national level (Table 1). This suggests that manufacturing employment growth could have a direct negative effect on district-level linguistic heterogeneity. Even as overall linguistic heterogeneity fell for India, average district-level linguistic heterogeneity increased by about 4.8 points between 1931 and 1961. I find that on average manufacturing employment growth held back this increase by a quite substantial 5.9 points.

My analysis shows that linguists' prediction of continued consolidation of languages appears be well founded in the case of India. The importance of communication for economic activity continues to grow, highlighted by India's recent outsourcing boom. It would not be surprising to see bilingualism continue to increase among speakers of minority languages, and for India's stock of languages decline substantially over the next century. The inherent in the decision to become bilingual points to an important role for language policy.

I describe aspects of mid-20th century Indian history relevant to my argument in section 2. I then present a simple model of bilingualism and language shift in section 3. After describing the data I collected in section 4, I develop my empirical approach and present OLS and IV results on the effect of manufacturing employment growth on bilingualism in section 5. Section 6 links manufacturing employment growth to changes in linguistic heterogeneity. Section 7 provides a summary and conclusion.

2 Mid 20th Century India

2.1 Structural Stagnation and Structural Change

India's economy was quite stagnant in the half century before 1920. Between the first all-India census in 1872 and the census of 1921, manufacturing consistently provided jobs to about 10% of the workforce. About 10% of the population lived in cities. There was virtually no growth in per-capita GDP over the 50-year period. India's stagnation broke during the 1920s and a period of structural change began. The economy began to shift employment into the manufacturing sector, and urbanization increased. Mortality also began to decline in the early 1920s, setting India's demographic transition into motion.

Between 1931 and 1961, India's structural transformation was rapid in comparison with the preceding half century, even if per-capita output did not grow much. Manufacturing employment grew at 2.7% annually, expanding from 7.4% to 11.1% of the total workforce. While this might not seem dramatic by modern standards, it is similar to the 3.1% annual growth U.S. manufacturing employment had between 1849 and 1879 when the industrial revolution took hold(Carter *et al.* 2006). India also became substantially more urban. In 1961, 18% of India's population lived in cities and towns, up from 12% in 1931.

Indian manufacturing enterprises increased substantially in scale between 1931 and 1961. Large factories, defined as those having more than 10 employees with power or 20 without power, provided 39.9% of all manufacturing jobs in 1961, more than double their 15.6% share in 1931. Historical studies have suggested that increased specialization was a major cause of the increase in scale (Roy 1999, 2000). The shift to larger work groups and greater task specialization increased the communication demands on workers. Labor productivity in large factories grew at a relatively brisk 2.1% annual rate between 1931 and 1947, while small factories actually saw a 1.5% annual decline in labor productivity (Sivasubramonian 2000). While both scale and the productivity of large factories were increasing, the bulk of Indian industry continued to use simple, labor intensive technologies (see Figure 2).

India's external environment and trade policy were important factors driving the structural shift toward manufacturing. India's exports in the early 20th century were primarily agricultural commodities such as tea, wheat, flax, raw cotton, and raw jute. The price of these export commodities relative to the manufactured goods India imported began fall in the late teens (Appleyard 2006). This negative terms of trade shock favored Indian manufacturing at the expense of agriculture. Additionally, in 1919, the government of India was given fiscal autonomy from Britain, which meant it could set tariff policy independently. At the same time, rights to land revenue were devolved to the provinces. Thereafter India's central government relied increasingly on import tariffs to raise revenue (Tomlinson 1979). Average import tariffs almost trebled from an average of 4.5% in the teens to 12.3% in the 1920s (Figure 3). Beginning in the mid 1920s, advocates for Indian industries successfully lobbied for protective tariffs. Average tariffs were 23.3% between 1931 and 1961, almost double the level of the 1920s.

2.2 Language, Literacy, and Education

India is one of the most linguistically diverse countries in the world. The probability that two randomly selected Indians share a mother tongue is only about 20%, similar to countries such as Nigeria, Kenya, and Indonesia. At least 180 distinct languages and about 600 dialects are indigenous to India. Dozens of these languages have literary traditions and a large minority of them are written using one of India's several scripts.

Although most of India's languages are concentrated in particular regions of the country, there is still substantial linguistic diversity within small geographic units (Table 2). The mother tongue of 23% of Indians was a minority language in their district of residence in 1931, rising to 26% in 1961. The average district has two or three minority languages with substantial population shares (Figure 4a).

The modern economy that India began to evolve during the late 19th and early 20th century created greater opportunities for people of different language groups to interact. Cities, modern industries, and the railways all brought people into contact. Figure 5 gives a tangible sense of the interaction between different language groups in public spaces during the middle 20th century.

Between 1931 and 1961, the average bilingualism rate among minority-language mother tongue speakers increased from 28.2% to 43.8% (Table 2). Bilingualism was negatively correlated with the size ranking of a language in its district in 1931. Most of the growth in bilingualism between 1931 and 1961 happened among speakers of medium size minority languages ranked 2 through 4 (Figure 4b). Nearly 80% of minority language mother tongue speakers who were bilinguals chose the majority language of their district as their second language, while the remainder generally chose either English or Hindi. Bilingualism also increased substantially among speakers of the local majority language.

Literacy was expanding rapidly in mid-20th century India. About 24.0% of adults could read in 1961, up from just 9.5% in 1931. Most languages that enter my analysis have written forms, so desire for literacy *per se* probably did not generate substantial second language acquisition. In fact, it is possible that literacy and bilingualism are substitutes. However, while formal schooling in the vernacular languages of India and in English had been promoted since the 1850s, the primary school completion rate was very low even in 1961. Many children attended primary school for a year or two, perhaps long enough to attain a basic literacy, but few finished. The Census of India did not ask about schooling until 1941 (Srivastava 1972); in 1961, only 7.0% of the population had completed the three to four years that comprised primary school.

3 The Economics of Bilingualism and Language Shift

I will now develop a parsimonious model that links the decision to acquire a second language to imperfect sorting in the labor market and higher productivity enabled by communication. This model is related to previous work by Lang (1986), who developed a language-based theory of discrimination, and Lazear (1999, 2005), who modeled the linguistic assimilation of immigrants.

As I argued in section 1, interaction with others allows individuals to take advantage of gains from trade of many different types, such as working together on a project, sharing information, or trading commodities. Knowledge of a second language expands the network of individuals with whom one can potentially interact. Gains from trade can thus provide individuals with incentives to expand their communication network by becoming bilingual. Imperfect sorting implies there will be an externality in language acquisition, as in the model of Church & King (1993). If I incur a cost to learn your language, some benefit may accrue to you from now being able to interact with me.The return to communication affects parents' decisions about which languages their children will learn. Bilingual parents can choose whether their children will learn one or both of their languages via the language they use in interacting with the children. Parents may discourage a child's acquisition of a language that only expands the child's communication network a small amount or which is socially stigmatized.

Consider a two-period economy with two production sectors, manufacturing and agriculture. The economy is populated by N dynasties. Each dynasty has one worker alive in period 0 and one alive in period 1. Workers are endowed with one unit of labor and engage in production in each period. Both sectors produce the same final good, the price of which is normalized to 1. Workers care about overall consumption for their dynasty j: $U_j = c_j^0 + c_j^1$.

Two languages are spoken in the economy. A majority of period 0 workers speak the dominant language D while a minority speak the secondary language S: $p_D^0 > \frac{1}{2} > p_S^0$. Some workers may be bilingual. The population shares of monolingual D and S speakers in period t are m_D^t and m_S^t; the share of bilinguals is b^t. These shares sum to one: $m_D^t + m_S^t + b^t = 1$. The period t population shares of everyone able to speak D and S, whether as monolinguals or as bilinguals, are $p_D^t = 1 - m_S^t$ and $p_S^t = 1 - m_D^t$.

The manufacturing sector makes use of a more productive technology than the agricultural sector. I take technology to be exogenous. The manufacturing technology requires workers to communicate with each other to take advantage of its superior productivity. This assumption reflects the need for manufacturing workers to coordinate their activities more intensively and across larger groups than agricultural workers. Manufacturing firms tend to be larger in scale and to rely on the labor market to find workers. Farmers tend to draw labor from the family.

At the beginning of each period, workers are randomly paired into firms. If members of a firm share a language in common, they are capable of jointly operating the manufacturing technology. Otherwise they are only capable of working in agriculture. Common-language firms get access to the manufacturing technology with the exogenous probability $\pi \epsilon (0, 1]$, which reflects how widespread manufacturing is in the economy. Workers in manufacturing each earn the return w_M. Workers in firms that do not share a common language or did not get access to the manufacturing technology must use the agricultural technology. Workers in agriculture each earn the return $w_A \leq w_M$. The expected period 0 income of a monolingual D speaker is $p_D^0(\pi w_M + (1-\pi)w_A) + (1-p_D^0)w_A$. A parallel expression holds for monolingual S speakers. Bilinguals earn $\pi w_M + (1-\pi)w_A$ since they can communicate with anyone with whom they are paired and always form a common-language firm. While workers in the real world target their job search based on where they think the opportunities are best rather than taking those that come randomly, this simple framework captures the intuitively appealing idea that there is at least some randomness in the matching process and that membership in a bigger network leads to better matches between worker and firm.

After workers are matched and produce in period 0, they give birth to one child and decide how much to invest in its language ability. Period 0 workers may costlessly transmit one of the languages they know to their child. Bilingual workers may transmit both languages by paying the cost $s_j \sim U[0, s]$, reflecting the additional effort parent and child must put in to disentangle the two languages. Monolingual period 0 workers may also invest in making their child bilingual by paying $c_j \sim U[0, c]$, which varies according to the quality of language instruction. The two costs are independent. Workers can costlessly borrow against period 1 income to finance investment if they wish. I will assume that N is large enough that workers do not find it possible to coordinate in making their investment decision. Once period 0 workers have made their investment decision, the period ends.

A monolingual S speaker will invest in bilingualism if doing so increases the expected income of his or her dynasty in period 1. This will be the case if the expected income from forming a common-language firm with certainty less the cost of bilingualism is greater than the expected income of a monolingual S speaker in period 1:

$$\pi w_M + (1-\pi)w_A - c_j \quad \geq \quad p_S^1(\pi w_M + (1-\pi)w_A) + (1-p_S^1)w_A. \tag{1}$$

A parallel inequality holds for monolingual D speakers. Define $\lambda = \pi(w_M - w_A)$ as the expected increase in return for a worker in a common language firm over a firm without a common language and recall that $m_D^1 = 1 - p_S^1$. Equation 1 can then be rewritten as

$$\lambda m_D^1 \geq c_j. \tag{2}$$

The benefit to a monolingual S speaker from becoming bilingual is the expected increase in return from forming a common language firm multiplied by the probability of matching with someone who only speaks D, in which case bilingualism would enable the formation of a common language firm. The shares q_S and q_D of monolingual S and D speakers for whom the benefits of becoming bilingual outweigh the costs are given by:

$$q_S = \begin{cases} \frac{\lambda}{c} m_D^1 & \text{if } m_D^1 < \frac{c}{\lambda}\text{, and} \\ 1 & \text{otherwise;} \end{cases} \tag{3}$$

$$q_D = \begin{cases} \frac{\lambda}{c} m_S^1 & \text{if } m_S^1 < \frac{c}{\lambda}\text{, and} \\ 1 & \text{otherwise.} \end{cases} \tag{4}$$

Bilinguals must decide whether to pass one or both languages to their children. Let the shares of bilinguals that assimilate to become monolingual S and D speakers be a_S and a_D. Assimilating bilinguals will always have higher expected earnings in period 1 if they speak D because $p_D^1 > p_S^1$. Therefore, no bilingual will want its child to become a monolingual S speaker and $a_S = 0$. A bilingual will decide make its child a monolingual D speaker if the expected additional return from being able to form a common language firm if matched with a monolingual S speaker in period 1 is less than the cost of transmitting S to the child: $\lambda m_S^1 \leq s_j$. This implies that:

$$a_D = \begin{cases} 1 - \frac{\lambda}{s} m_S^1 & \text{if } m_S^1 < \frac{s}{\lambda}\text{, and} \\ 1 & \text{otherwise.} \end{cases} \tag{5}$$

Workers make the investment decision at the end of period 0 anticipating the equilibrium share of workers that will be able to speak S and D in period 1. The period 1 population shares that speak S and D in turn depend on the decisions of the monolingual workers in period 0:

$$p_S^1 = p_S^0 + q_D m_D^0 - a_D b^0 \tag{6}$$
$$p_D^1 = p_D^0 + q_S m_S^0. \tag{7}$$

I use equations 3 to 7 to solve for the equilibrium shares q_S and q_D in terms of the initial distribution of language ability. I assume that $q_S, q_D, a_D < 1$.

$$q_S = \frac{\frac{\lambda}{c}}{1 - \frac{\lambda^2}{c^2} m_D^0 m_S^0 - \frac{\lambda^2}{cs} m_S^0 b^0} \left(m_D^0 \left(1 - \frac{\lambda}{c} m_S^0\right) + b^0 \left(1 - \frac{\lambda}{s} m_S^0\right) \right) \tag{8}$$

$$q_D = \frac{\frac{\lambda}{c}}{1 - \frac{\lambda^2}{c^2} m_D^0 m_S^0 - \frac{\lambda^2}{cs} m_S^0 b^0} \left(m_S^0 \left(1 - \frac{\lambda}{c}(m_D^0 + b^0)\right) \right). \tag{9}$$

The shares will be strictly between 0 and 1 as long as the expected wage in a common-language firm is not too large relative to the cost of becoming bilingual or transmitting two languages. I assume that λ is sufficiently less than c and s for this to be the case.

3.1 Language Share, Manufacturing, and Bilingualism

Three results flow from equations 8 and 9 that link bilingualism to the population share speaking each mother tongue and to the expected wage in a common-language firm.

Result 1 *A larger share of monolingual speakers will become bilingual when they are a smaller share of the population.*

This follows directly from differentiating q_S and q_D with respect to the initial monolingual population shares: $\frac{\partial q_S}{\partial m_S^0} < 0$ and $\frac{\partial q_D}{\partial m_D^0} < 0$.

Result 2 *A larger share monolinguals will become bilingual when the return to being in a common-language firm is greater.*

This follows from differentiating q_S and q_D with respect to λ. Under the assumptions that $m_S^0 < \frac{1}{2}$, $\lambda < c$ and $\lambda < s$, we have $\frac{\partial q_S}{\partial \lambda} > 0$. The sign of $\frac{\partial q_D}{\partial \lambda}$ depends on parameters. If $m_D^0 + b^0 < \frac{c}{2\lambda}$, then $\frac{\partial q_D}{\partial \lambda} > 0$.

Result 3 *The incentive to become bilingual generated by the return to being in a common-language firm is larger for the minority S speakers than the majority D speakers.*

Differentiating q_S and q_D as before and using the assumption that $p_S^0 < p_D^0$ gives the result that $\frac{\partial q_S}{\partial \lambda} > \frac{\partial q_D}{\partial \lambda}$. Intuitively, because D speakers are a larger share of the population, the additional return from being able to form a common-language firm with certainty is lower, while the cost c_j of becoming bilingual is fixed. D speakers will thus have a lower communication-based incentive to learn S than S speakers will have to learn D.

3.2 Bilingualism and Language Shift

Language shift within a lineage results when a bilingual parent transmits only their second language to their children. The share of the population knowing S will decline if the number of bilinguals in period 0 who are better off assimilating is greater than the number of D monolinguals who find it worthwhile to learn S.

$$\Delta p_S = q_D m_D^0 - a_D b^0 \tag{10}$$

Whether Δp_S is positive or negative depends on the parameters.

Result 4 *Bilingualism among speakers of S may lead to a decline over time in the share of the population speaking S.*

Bilingual dynasties may remain bilingual, even if one of their languages has very few or no monolingual speakers and therefore low communication value. There are examples of stable bilingualism of this type. In Wales, for example, about 20% of the population currently speaks Welsh even though there are no longer any monolingual Welsh speakers. Linguists believe the socio-cultural value of language is an important factor in whether the children of bilingual parents are raised as monolinguals. For example, if members of a linguistic minority face discrimination, they may be more likely to assimilate. Alternatively, if a minority language is culturally valorized, as in the case of Welsh, assimilation may be less likely. My model does not directly account for this factor, but doing so would not change result 4.

The economics literature commonly uses a Herfindahl index h to measure of linguistic heterogeneity: $h = 1 - \sum_\ell s_\ell^2$, where s_ℓ^2 is the population share of speakers whose mother tongue is ℓ. When everyone has the same mother tongue $h = 1$ and when nobody does $h \rightarrow 0$. Result 4 implies that the growth of manufacturing employment may lead to a decline in linguistic heterogeneity.

4 Panel Data for Indian Districts

I constructed a panel dataset of Indian districts for the years 1931 and 1961 from the Census of India to test the intuitions developed in the model (India 1933, 1962). Sample data for North Arcot district illustrates the data structure (Table 4). North Arcot is a large district and had about 2.3 million inhabitants in 1931[1]. It is located about 200 km east south-east of Madras. Panel A shows district-level variables x_{dt} while panel B shows district × language variables $x_{\ell dt}$.

Tamil is the majority language of the district. It was the mother tongue of 80% of the population in 1931 and 82% in 1961. The five largest minority languages in 1931 were Telugu, Hindi, Kannada, Saurashtri, and Malayalam. Bilingualism is very prevalent among minority language speakers. In contrast to the average district, bilingualism in North Arcot generally declines over time. North Arcot is a little less linguistically heterogeneous than the average district, and becomes less heterogeneous over time.

I assembled data in this form for all districts in 1931 and 1961. I first selected the six most commonly spoken languages in 1931. I collected the number of speakers and bilinguals

[1] North Arcot is composed of the present-day districts of Tiruvannamalai and Vellore.

of these languages for both 1931 and 1961. I also collected information on employment, urbanization, and literacy for each district and year. In each year, each district has six language-level observations and one observation of employment and other district-level characteristics. The dataset contains 153 districts and covers all of present-day India excluding Uttar Pradesh, Punjab, Himachal Pradesh, and Rajasthan. See Appendix A for further details on how the dataset was constructed.

5 Effects of Manufacturing Growth on Bilingualism

I have argued that bilingualism is a necessary condition for language shift. The first three predictions I derived in section 3 link relative language size and manufacturing employment to bilingualism. I test these predictions by estimating an econometric model in differences using my 1931–1961 district-level panel data set. Differencing removes a fixed effect for each district-language combination.

$$\Delta b_{\ell d} = \beta_c + \beta_{min} I_{\ell d}^{min} + \beta_m \Delta m_d + \beta_m^{min} (\Delta m_d \times I_{\ell d}^{min}) + D_d' \Theta + L_{\ell d}' \Psi + \Delta \varepsilon_{\ell d}. \tag{11}$$

The dependent variable $\Delta b_{\ell d}$ is the change in the share of mother tongue speakers of ℓ in district d that are bilingual. The indicator $I_{\ell d}^{min}$ takes on the value 1 if language ℓ is a minority language in district d, which I take to be exogenous.[2] Change in the manufacturing share of employment is Δm_d. The vector D_d' contains additional district-level controls, including changes in urbanization, the literacy rate, the employment rate, and initial levels of variables. The vector $L_{\ell d}'$ contains language × district controls, such as the population share speaking language ℓ in district d.

The dependent variable $\Delta b_{\ell d}$ describes the change from 1931 to 1961 in the stock of bilinguals for each mother tongue ℓ in each district d. It is closely related to the transition probabilities q_S and q_D from the model. The stock of bilinguals increases when mother tongue speakers of ℓ become bilingual. Both the stock of bilinguals and the stock of mother tongue speakers of ℓ decline when a bilingual person assimilates to his or her second language. Therefore, when a lineage undergoes the transition from monolingualism to bilingualism, $\Delta b_{\ell d}$ goes up. When a lineage assimilates, $\Delta b_{\ell d}$ goes down but by a smaller amount. We see large increases in $b_{\ell d}$ over the panel, suggesting assimilation is relatively slow.

The model predicts that manufacturing employment will have a positive effect on bilingualism, suggesting the coefficients β_m and β_m^{min} will be positive and that $\beta_m^{min} > \beta_m$. Bilingualism should be more prevalent among speakers of smaller languages if sorting is imperfect,

[2] There was no change in the majority language of any district over the panel.

so I expect the coefficient on the control for change in a language's population share to be negative.

When I estimate equation 11, I weight each observation by the number of speakers of language ℓ in district d in 1931. This allows me to interpret the coefficients as effects on an average individual. Because the variables of interest are at the district level and the language data is at the language-district level, I allow for arbitrary correlation of $\Delta\varepsilon_{\ell d}$ at the district level when computing the standard errors.

I also include a district-specific trend v_d in some estimations. District-level variables are absorbed by the trend and are not included separately in this specification.

$$\Delta b_{\ell d} = \beta_c + \beta_{min} I_{\ell d}^{min} + \beta_m^{min}(\Delta m_d \times I_{\ell d}^{min}) + L_{\ell d}'\Psi + v_d + \Delta\xi_{\ell d}. \tag{12}$$

Of the effects that interest, equation 12 can identify only β_m^{min}, the differential effect of manufacturing employment growth for minority language mother tongue speakers; the level effect of manufacturing employment growth is absorbed by the trend.

All estimations control for the 1931 levels of urbanization, literacy, and the workforce participation rate. OLS estimates will nevertheless suffer from reverse causality and omitted variables bias. I will first examine baseline OLS estimations of equations 11 and 12. I will then offer a discussion of bias and an instrumental variables solution to it. I then present IV estimates.

5.1 Baseline OLS estimates

OLS estimation of equation 11 is consistent with the main predictions of the model. Table 5 shows that a one point increase in the manufacturing share of employment is correlated with statistically significant increases in the bilingual share of 0.11 for speakers of majority languages (column 1). The effect for minority languages is an additional 1.10 points for a total effect of 1.22 points. These estimates are consistent with result 3 of the model, which said that the effect of manufacturing employment should be larger for minority languages than for the majority language. The absolute effect of manufacturing employment on bilingualism is quite big; 1.6 additional manufacturing jobs in a district results in one additional bilingual for both majority and minority language speakers. Controlling for the initial levels of urbanization, literacy, employment, and language size do not change the estimates much (Table 5, column 2). The positive difference between the manufacturing employment effect for minority and majority languages is attenuated to 0.76 points when I introduce a district-specific trend into the estimation, though it is still strongly significant (Table 5, column 3). The basic pattern is robust to estimation of the model in levels with only district fixed effects or with no fixed effects at all.

5.2 Sources of Reverse Causality and Omitted Variables Bias

OLS estimation of equation 11 likely suffers from both simultaneity and omitted variable bias. Simultaneity bias can result from the effect of language on economic outcomes, which is discussed in the literature. There are several possible sources of omitted variable bias, including migration and education. In addition, the census measure of literacy I use as a control is noisy, and may leave some residual bias.

A possible mechanism of reverse causation is that bilingualism makes communication easier, and therefore districts with a growing bilingual share may attract more manufacturing firms. Alternatively, bilingualism may be correlated with other forms of human or social capital that is attractive to firms, such as the absence of ethnic conflict. These relationships will bias the OLS estimates of β_m and β_m^{min}. This bias could be either positive or negative.

Migration may be induced by a positive economic shock to a district. Migration might be away from a district, because income aids mobility, or into a district, because it is doing well. A positive shock from the discovery of a natural resource, for example, could aid manufacturing, while climatic shocks favoring agriculture might hinder it by increasing wages. Bilingualism is a human capital investment, and owing to credit constraints residents of districts with positive economic shocks invest more. Migrants may also be positively selected for bilingualism for reasons besides their wealth. This bias could also go either up or down, depending on whether inmigration of bilinguals outweighs outmigration and whether the shock aids or hinders manufacturing.

Education is offered in a limited set of languages in most countries, both for efficiency and to encourage social integration. India is no exception, but completion of even primary school was uncommon in this period. Nevertheless, the demand for education doubtless encouraged some minority-language speakers to become bilingual. Educated workers may attract manufacturing firms and may themselves be more likely to move to the city, producing an upward bias in my estimates. Alternatively, manufacturing may make a district richer and increase spending on education, encouraging bilingualism in a different way that also leads to a positive bias.

5.3 Instrument Based on the Manufacturing Mix

I construct an instrumental variable for the change in manufacturing employment to mitigate simultaneity and omitted variables bias. My instrumental variable is a prediction of how fast manufacturing employment in each district would have grown if each of its subindustries had grown at the average rate for the rest of the country. It therefore isolates the component of district-level manufacturing employment growth that resulted from national-level variation in employment demand by subindustry. These factors include final demand, relative productivity growth, tariffs, and world prices.

I link national changes to the district level via the initial industrial mix within manufacturing for each district. The mix of subindustries located in a particular district is in part due to enduring district attributes, such as its location relative to raw materials sources or major markets. A similar approach has been used to explore the effect of employment shocks on city-level economic variables in the United States (Blanchard & Katz 1992; Bartik 1991).

To create the instrument, I collected data on 1931 manufacturing employment in nine subindustries: textiles, wood, metals, ceramics, chemicals, apparel, food processing, vehicles, power, and other. I first estimate regressions of the change in the manufacturing share of the workforce on the difference in initial subindustry workforce shares from the subindustry means.

$$\Delta m_d = \psi + \sum_i \mu_i (y_{id} - \overline{y_i}) + \zeta_{id}. \tag{13}$$

The y_{id} are the 1931 shares of the workforce in manufacturing subindustry i in district d. The average share of the district workforce in industry i is $\overline{y_i}$. The coefficient ψ measures the average change in manufacturing employment between 1931 and 1961. The coefficients μ_i measure the effect on manufacturing employment growth of having higher- or lower-than-average shares of the workforce in each subindustry. I then use the coefficients from that regression and the mean-deviated subindustry shares for district j to make an out-of-sample prediction $\widehat{\Delta m_j}$.

The exclusion restriction for this instrument is that initial subindustry shares y_{id} and the mean-deviated share growth coefficients μ_i are uncorrelated with the unobserved determinants of the change in bilingualism in equations 11 and 12.

States have industrial specialization, and it is possible that some industries are sufficiently concentrated that state-level policy could influence the national growth of a particular industry. This could lead μ_i to be correlated with the unobservables. To eliminate this problem, I exclude those districts that share a state or province with j in either 1931 and 1961 when I estimate 13.

There are scenarios under which the initial share of workers in a particular subindustry is related to the change in bilingualism through some channel other than Δm_d. It seems unlikely these scenarios would result in a large effect, and since the instrument is strong as shown below, the likely bias will be small. The following example illustrates a potential criticism. Suppose a district begins with a large share of workers in a subindustry that offers year-round employment. Year-round employment may make it more likely a worker will migrate with his or her family. National final demand growth in that subindustry may induce relatively more migration of complete families from outside the district. The non-manufacturing workers in those families may be positively selected for bilingualism. In such a case, the exclusion restriction will be violated. In estimations where I include a district-specific trend, this channel would have to affect majority and minority language speakers differentially to violate the exclusion restriction.

Men and women tend to work in different types of firms within a given industrial classification. To improve the instrument's power, I include male and female workforce shares separately when constructing it. We can see the instrument at work in a sample estimation that includes all districts (see Table 6). The estimation has an R^2 of 0.69, showing that the initial industrial shares are very good at predicting where manufacturing will expand or contract. Recall that broad tariff increases were an important factor driving manufacturing expansion. In the textile sector, male employment is associated with growth of the manufacturing employment share, while female employment is associated with its decline. Women's textile jobs in 1931 were more heavily biased toward home weaving and spinning, which declined relative to factory weaving over the period.

I estimate the first stage of an IV estimation of equation 11 by replacing the bilingual share $\Delta b_{\ell d}$ with the endogenous variable of interest and by including the instruments $\widehat{\Delta m}_d$ and $\widehat{\Delta m}_d \times I_{\ell d}^{min}$ as regressors. Recall $I_{\ell d}^{min}$, the indicator for whether language ℓ is a minority language in district d, is exogenous. The instrument strongly predicts both the change in the district-level manufacturing share and its interaction with the minority language indicator in the standard estimation and the estimation that includes district level trends (Table 7). F-tests of the excluded instrument are larger than the critical values that would indicate a weak instrument (Staiger & Stock 1997; Stock & Yogo 2002).

5.4 IV Estimates of the Effect of Manufacturing Employment on Bilingualism

Instrumental variables estimates of the manufacturing employment effect are similar to the OLS estimates (Table 8). A one point change in the manufacturing share of employment leads to a 1.3 point increase in bilingualism among minority language speakers and a 0.37 point increase in bilingualism among majority language speakers. These point estimates are statistically significant at 1%. As the model suggested, the effect of increased manufacturing employment is to encourage bilingualism, particularly for speakers of minority languages. The absolute effect of manufacturing employment on bilingualism remains big; 1.6 additional manufacturing jobs in a district results in one additional bilingual for both majority and minority language speakers.

How much of the changes in bilingualism between 1931 and 1961 can my instrumental variables estimates explain? Bilingualism increased by an average 15.6 percentage points among minority language speakers and an average 5.1 percentage points among majority language speakers. The population-weighted average manufacturing share of employment increased by 4.2 points between 1931 and 1961 (Table 2). These estimates suggest that manufacturing employment growth accounts for about a third of the the increase in bilingualism among both minority and majority language speakers.

5.5 Cultivators and Agricultural Laborers

I motivated the idea that manufacturing employment would be associated with increased bilingualism by contrasting the role of communication in manufacturing and agriculture. I argued that the specialization and scale that generate productivity gains in manufacturing require more intensive and extensive coordination of tasks between workers. If such forces are underlying the positive effect of the manufacturing employment share on bilingualism, the agricultural share of employment should have a negative sign when substituted for the manufacturing employment share in an OLS estimation of equation 11. If the positive coefficient on manufacturing was driven by a shift of workers from commerce to manufacturing, for example, my motivation would be much less convincing. I find the agricultural share of employment indeed has a strong and significant negative effect on minority language bilingualism and a smaller effect for majority languages (Table 9, column 1). Estimation with district trends shows the effect on minority languages to be more negative than the effect for majority languages (Table 9, column 2).

Most agricultural workers in India are owner-operators or tenants. The remainder are agricultural laborers. A further check on the consistency of my motivation is to look within the agricultural workforce at the share who are agricultural laborers. Agricultural laborers rely on the labor market to find work and engage in temporary employment; owner-operators and tenants are much less mobile. It would be more advantageous *ceteris paribus* for agricultural laborers to be able to communicate widely than owner-operators and tenants. I replace the manufacturing share in equation 11 with the share of agricultural workers who are agricultural laborers are re-estimate the model (Table 9, panel B, column 1). I find that the agricultural labor share of the agricultural workforce has a positive effect on bilingualism for minority language speakers, which is consistent with the view that the communication intensity of different sectors drives their effect on bilingualism. An estimation including district trends shows the effect on minority languages to be more positive than the effect for majority languages and is statistically significant at 10% (Table 9, panel B, column 2).

6 Manufacturing Employment and Linguistic Heterogeneity

It is possible that the two channels I have discussed so far are uncoupled. While bilingualism is correlated with language shift (Table 1, it is possible that bilingualism resulting specifically from manufacturing employment growth is not. For example, it could be that manufacturing employment encourages minority speakers to obtain too shallow a knowledge of the majority language to allow them to make it the sole language of the next generation, while bilingualism that results from other causes results in a deeper knowledge.

If large languages grow in relative size at the expense of smaller ones, linguistic heterogeneity will fall. This is the outcome we also expect from language shift. The effect of effect of manufacturing employment growth on linguistic heterogeneity sheds light on whether the two channels are coupled and is also of independent interest as it has been associated with poor economic performance.

Linguistic heterogeneity declined between 1931 and 1961 in the area covered by the panel, falling from 0.87 to 0.84.[3] Average district-level linguistic heterogeneity actually increased from 0.30 to 0.35, meaning that languages were becoming less geographically concentrated even as they were consolidating at the national level.

I estimate the effect using a version of equation 11 in which there are only district-level regressors. IV estimates show a one point increase in the manufacturing share of employment leads to a statistically significant 1.50 point decrease in district-level linguistic heterogeneity (Table 10, column 3). IV estimates are close to the OLS. District-level linguistic heterogeneity increased in India between 1931 and 1961 from 0.30 to 0.35 (Table 2). Part of this increase is likely due to migration across districts with different majority languages. Manufacturing employment growth actually slowed this trend; in its absence mean linguistic heterogeneity would have climbed to 0.41 by 1961.

It is difficult to form a precise comparison between my district-level results and estimates from the literature of the economic effects of linguistic heterogeneity. The economic variables and the units across which heterogeneity is measured differ. The channels through which linguistic diversity affects economic outcomes may be different at the country and local levels. My estimates imply a one standard deviation increase in manufacturing employment

[3] Calculated as $1 - \sum_\ell s_\ell^2$, where s_ℓ is the share of the population speaking ℓ.

leads to a 0.30 standard deviation decrease in linguistic heterogeneity. The best cross-country estimate from Alesina & La Ferrara (2005) is that a one standard deviation increase linguistic heterogeneity leads to 0.6% lower annual per-capita GDP growth. They find that county-level heterogeneity in the United States has no effect on population growth, which they use as a proxy for economic growth. However, Alesina *et al.* (1999) found that a one standard deviation increase in racial heterogeneity in U.S. counties lead to 0.25 standard deviation lower spending on roads. My estimates linking economic change to linguistic heterogeneity are of roughly similar magnitude to those in the literature linking heterogeneity to economic variables.

7 Conclusion

I have shown that manufacturing employment growth leads to growing bilingualism among speakers of local minority languages, and that bilingualism is associated with the relative decline of a mother tongue. Manufacturing employment growth discourages linguistic heterogeneity. Part of the measured negative impact of linguistic heterogeneity on economic growth and public goods will be confounded by the process of economic development itself. My results suggest that language learning and linguistic diversity ought to be taken as endogenous to the process of economic development. Further, they suggest that the expansion of economic activities in which language is an important driver of linguistic consolidation.

The importance of communication for economic activity continues to grow. My analysis shows that linguists' prediction of continued consolidation of languages worldwide appear be well founded, particularly for India. India has seen rapid growth of manufacturing and services in recent decades, and it would not be surprising to see India's stock of languages decline to a few dozen in the next century. This means a large number of monolingual Indians will become bilingual and that many bilinguals will raise monolingual children.

It is reasonable to suppose that, as with with other networks, the benefit of knowing a language is increasing in its size, giving rise to a network externality. This externality will be positive when someone decides to become bilingual and negative when a bilingual decides to raise its children as monolinguals. India's linguistic transition is thus likely to proceed more slowly than would be optimal, as those who pay the cost of bilingualism create benefits for other speakers of the languages they learn. Assimilation has the opposite effect; a parent's decision to abandon a language reduces the size of the language in the next generation and negatively affects the remaining speakers.

Whether or not there are economically important network externalities in language ability needs to be demonstrated empirically. If these externalities exist and are substantial, there may be a significant role for government in transferring resources to minority language speakers to encourage bilingualism and to ease the negative effects of assimilation on those who remain monolingual linguistic minorities.

References

Alesina, Alberto, & La Ferrara, Eliana. 2005. Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, **42**, 762–800.

Alesina, Alberto, Baqir, Reza, & Easterly, William. 1999. Public Goods and Ethnic Divisions. *Quarterly Journal of Economics*, **114**(4), 1243–84.

Anderson, James, & van Wincoop, Eric. 2004. Trade Costs. *Journal of Economic Literature*, **XLII**, 691–751.

Appleyard, Dennis R. 2006. The Terms of Trade between the United Kingdom and British India, 1858-1947. *Economic Development and Cultural Change*, **54**(3), 635–655.

Bartik, Timothy J. 1991. *Who Benefits from State and Local Economic Development Policies?* W.E. Upjohn Institute.

Berman, Eli, Lang, Kevin, & Siniver, Erez. 2003. A Language Theory of Discrimination. *Labor Economics*, **10**, 265–290.

Blanchard, Olivier, & Katz, Lawrence. 1992. Regional Evolutions. *Brookings Papers on Economic Activity.*

Bleakley, Hoyt, & Chin, Aimee. 2004. Language Skills and Earnings: Evidence from Childhood Immigrants. *Review of Economics and Statistics*, **86**(2), 481–496.

Carter, Susan B., Gartner, Scott Sigmund, Haines, Michael R., Olmstead, Alan L., Sutch, Richard, & Wright, Gavin (eds). 2006. *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition.* Cambridge University Press.

Chiswick, Barry, & Miller, Paul. 1995. The Endogeneity between Language and Earnings: International Analyses. *Journal of Labor Economics*, **13**(2), 246–288.

Church, Jeffrey, & King, Ian. 1993. Bilingualism and Network Externalities. *The Canadian Journal of Economics / Revue Canadienne d'Economique*, **26**(2), 337–345.

Crystal, David. 1997. *Language Death.* West Nyack, NY: Cambridge University Press.

Dustmann, Christian, & van Soest, Arthur. 2001. Language fluency and earnings: estimation with misclassified language indicators. *Review of Economics and Statistics*, **83**, 663674.

Fishman, Joshua. 1964. Language Maintenance and Language Shift as a Field of Inquiry. *Linguistics*, **9**, 32–70.

Gal, Susan. 1978. *Language Shift: Social Determinants of Language Change in Bilingual Austria.* New York: Academic Press.

Gordon, Raymond G. 2005. *Ethnologue: Languages of the World.* 15th edn. Dallas, Tex: SIL International.

Hensley, Glenn S. 1944. *Sawing Logs*. Photograph held by the University of Chicago Library: http://dsal.uchicago.edu/images/hensley.

Hill, Jane H. 1978. Language Death, Language Contact, and Language Evolution. *In:* McCormack, William C., & Wurm, Stephen A. (eds), *Approaches to Language: Anthropological Issues*. The Hague: Mouton.

India. 1933. *Census of India 1931*. Census Commissioner, Government of India.

India. 1962. *Census of India 1961*. Census Commissioner, Government of India.

India. 1991. *Census of India 1991*. Census Commissioner, Government of India.

Krauss, Michael. 1992. The World's Languages in Crisis. *Language*, **68**(1), 4–10.

Lang, Kevin. 1986. A Language Theory of Discrimination. *Quarterly Journal of Economics*, **101**(2), 363–382.

Lazear, Edward. 1999. Culture and Language. *Journal of Political Economy*, **107**(6), S95–S126.

Lazear, Edward. 2005. *The Slow Assimilation of Mexicans in the United States*. Unpublished Ms.

Roy, Tirthankar. 1999. *Traditional Industry in the Economy of Colonial India*. Cambridge: Cambridge University Press.

Roy, Tirthankar. 2000. *The Economic History of India, 1857-1947*. Oxford University Press.

Singh, R.P., & Banthia, Jayant Kumar. 2004. *India Administrative Atlas, 1872-2001: A Historical Perspective of Evolution of Districts and States in India*. New Delhi: Controller of Publications.

Sivasubramonian, S. 2000. *The National Income of India in the Twentieth Century*. New Delhi: Oxford University Press.

Srivastava, Shyam Chandra. 1972. *Indian Census in Perspective*. New Delhi: Office of the Registrar General.

Staiger, D., & Stock, J.H. 1997. Instrumental Variable Regression with Weak Instruments. *Econometrica*, **65**(3), 557–586.

Stock, James H., & Yogo, Motohiro. 2002. *Testing for Weak Instruments in Linear IV Regression*. NBER Working Paper No. T0284.

Tomlinson, B.R. 1979. *The Political Economy of the Raj, 1914-1947 : The Economics of Decolonization in India*. London: Macmillan Press.

Weber, Eugen. 1976. *Peasants into Frenchmen : the Modernization of Rural France, 1870-1914*. Palo Alto: Stanford University Press.

Appendix

A Construction of the Datasets

I constructed a two panel datasets using the Census of India (India 1933, 1962, 1991). The first panel includes Indian districts for the years 1931 and 1961. Many district boundaries changed following India's independence from Britain in 1947, when hundreds of sovereign princely states were integrated into the colonial administrative framework inherited by India. Some British districts were also combined or split up. The census does not always contain sufficient detail in 1961 to use the 1931 district definitions in constructing the dataset. I aggregated geographical units as necessary to form exactly comparable districts based on the equivalence table found in Singh & Banthia (2004). The aggregation produced 244 comparable districts, compared with 339 administrative districts in 1961 and 439 administrative districts, princely states, and territories in 1931. Only some of these aggregate districts are included in the dataset because 1931 bilingualism data does not exist for the states of Uttar Pradesh, Rajasthan, Punjab and Himachal Pradesh; local officials in the various provinces controlled the Census tabulation before independence and did not always produce the same tabulations. I exclude these states from the dataset, leaving 153 districts. For each district and year, I compiled characteristics of the six languages most commonly spoken in 1931. I thus have six language-level observations per district per year, and one observation of employment and other district characteristics per district per year.

I also constructed a panel dataset at the state level for 1961 and 1991. This dataset has a similar structure. There are 23 states in the dataset, each of which has information on up to 56 distinct languages.

Table 1: Initial Bilingualism and Language Consolidation

	Growth of Speakers	
Unit of Observation	State × Language	Language
Log Bilinguals 1961	-0.283**	-0.314***
	(0.08)	(0.12)
Log Speakers 1961	0.081	0.249**
	(0.08)	(0.10)
Constant	2.37***	1.03**
	(0.09)	(0.49)
State Fixed Effects	Yes	No
R^2	0.36	0.10
N	558	56

Notes: Standard errors are clustered at the state level and robust to heteroskedasticity at the language level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 2: Population-Weighted District-Level Summary Characteristics for 153 District Dataset

	1931	1961	Change 1931–1961
Employment Rate	0.461	0.439	-0.022
Manufacturing Share of Emp.	0.074	0.116	0.042
Urban Share	0.118	0.193	0.075
Bilingual Share of Population	0.069	0.096	0.027
Linguistic Heterogeneity	0.299	0.347	0.048
Literate Share of Population	0.069	0.261	0.192
Mother Tongue			
Speakers of Majority Language			
Population Share	0.773	0.739	-0.034
Share Bilingual	0.015	0.067	0.052
Mother Tongue			
Speakers of Minority Languages			
Population Share	0.227	0.261	0.034
Bilingual Share	0.282	0.438	0.156

Notes: Source is Census of India. Employment rate is the population-weighted district average share of the population who are in the workforce. Urban share is the population-weighted district average share of the population who live in urban areas. Bilingual share of the population is the population weighted average share of the district population who are bilinguals. Linguistic heterogeneity is population-weighted average of the district-level measure of linguistic heterogeneity $h_d = 1 - \sum_{\ell \epsilon d} s_{\ell d}^2$, where $s_{\ell d}$ is the population share of speakers whose mother tongue is ℓ in district d. Population share of speakers of the majority language is the population-weighted average district population share that speaks the majority language of the district. Share bilingual for majority language speakers is the population-weighted average share of speakers of majority languages in a district who are bilingual. Population share of speakers of minority languages is the population-weighted average district population share that speaks a minority language of the district. Share bilingual for minority language speakers is the population-weighted average share of speakers of minority languages in a district who are bilingual.

Table 3: All-India Industrial Mix in 1931 and 1961

	Overall		Males		Females	
	1931	1961	1931	1961	1931	1961
Manufacturing Share of Workforce	0.077	0.093	0.081	0.099	0.075	0.077
Share of						
Manufacturing Workers in						
Textiles	0.345	0.311	0.299	0.245	0.439	0.445
Wood	0.131	0.123	0.154	0.125	0.097	0.128
Metals	0.060	0.068	0.078	0.089	0.015	0.014
Ceramics	0.086	0.070	0.086	0.069	0.083	0.079
Chemicals	0.053	0.036	0.047	0.019	0.057	0.011
Apparel	0.142	0.110	0.168	0.136	0.081	0.047
Food Processing	0.119	0.165	0.083	0.165	0.215	0.254
Vehicles	0.002	0.033	0.003	0.045	0.000	0.001
Power	0.002	0.013	0.003	0.017	0.001	0.001
Other	0.060	0.071	0.079	0.091	0.012	0.019

Notes: From the Census of India. Equivalent subindustries for 1931 and 1961 were created using the 1901–1961 mapping of occupation codes in the 1961 census. This table includes workers in all states of post-Independence India and is thus not strictly comparable to Table 2.

Table 4: North Arcot: One of the 153 Districts in the Dataset

Panel A: District Characteristics

	Year	
	1931	1961
Population	2,266,989	3,146,326
Majority Language	Tamil	Tamil
Linguistic Heterogeneity	0.331	0.319
Workforce Share of Pop.	0.517	0.467
Manufacturing Share of Workforce	0.057	0.113
Urbanization Rate	0.149	0.201

Panel B: District-Level Language Characteristics by 1931 Language Rank

	Mother Tongue Share of Population		Bilingual Share of Mother Tongue	
	1931	1961	1931	1961
Tamil	0.800	0.815	0.017	0.040
Telugu	0.140	0.111	0.661	0.626
Hindi	0.042	0.001	0.643	0.562
Kannada	0.009	0.008	0.871	0.741
Saurashtri	0.002	0.001	0.663	0.649
Malayalam	0.000	0.001	0.650	0.688

Notes: North Arcot is a district of Madras State approximately 200 km east of the city of Madras. Linguistic heterogeneity is $h_d = 1 - \sum_{\ell \epsilon d} s_{\ell d}^2$, where $s_{\ell d}$ is the population share of the district whose mother tongue is ℓ.

Table 5: The Effects of Manufacturing Employment on Bilingualism: Weighted OLS Estimates in Differences

	Δ Bilingual Share of Speakers	
	(1)	(2)
Δ Manufacturing Share of Emp.	0.114**	
	(0.05)	
Δ Manufacturing Share of Emp.	1.103***	0.756***
× Minority Language	(0.28)	(0.24)
Minority Language	0.045	0.025
	(0.03)	(0.03)
Constant	-0.024**	0.045***
	(0.01)	(0.00)
District Trend Effects	No	Yes
R^2	0.649	0.589
N	824	824
Effects Computed from Coefficients		
Δ Manufacturing Share for	1.217***	
Minority Language Speakers	(0.30)	

Notes: Observations are at the language-district level and are weighted by number of speakers in 1931. Bilingual share of speakers is the ratio between the number of mother tongue speakers of language ℓ in district d who can speak a second language to the total number of mother tongue speakers of language ℓ in district d. Regressions include controls for initial levels of urbanization, literacy, workforce share of population, and language share of population. Standard errors are corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 6: Instrument using 1931 Employment Shares in Manufacturing Subindustries by Gender: Sample Regression

	Δ Mfg. Share		Δ Mfg. Share
Textiles	0.480***	Textiles × Female	-0.852***
	(0.13)		(0.15)
Apparel	-0.193	Apparel × Female	-0.337
	(0.29)		(0.92)
Wood	-1.128*	Wood × Female	2.146
	(0.48)		(1.60)
Metal	-1.569*	Metals × Female	-7.827***
	(0.72)		(2.16)
Ceramics	0.750	Ceramics × Female	-1.982
	(0.91)		(1.74)
Chemicals	-0.577	Chemicals × Female	-2.380
	(0.71)		(1.27)
Food	0.266	Food × Female	0.563
	(0.55)		(0.63)
Vehicles	27.528***	Vehicles × Female	-8.467
	(5.97)		(23.86)
Power	1.311	Power × Female	-11.052***
	(1.35)		(2.59)
Other	1.271*	Other × Female	4.884
	(0.54)		(3.07)
Constant	0.030***	Female	-0.002
	(0.01)		(0.01)
R^2	0.299		
N	153		

Notes: Observations at the district level. Standard errors corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 7: First Stage for Industrial Mix Instrument: Weighted OLS Estimates in Differences

	Δ Mfg. Share	Δ Mfg. Share × Minority Language	
	(1)	(2)	(3)
Predicted Δ Mfg. Share of Emp.	0.655***	-0.087***	
	(0.04)	(0.01)	
Predicted Δ Mfg. Share of Emp.	-0.495***	0.654***	0.647***
× Minority Language	(0.06)	(0.06)	(0.06)
Joint F-Test of Instruments	*177.22*	*63.38*	*126.18*
Shea's Partial R^2	*0.79*	*0.84*	*0.84*
Δ Mfg. Share of Emp.		0.140***	
		(0.02)	
Δ Mfg. Share of Emp.	0.744***		
× Minority Language	(0.09)		
Minority Language	-0.016***	0.023***	0.023***
	(0.00)	(0.01)	(0.01)
Constant	0.013***	-0.003***	0.000
	(0.00)	(0.00)	(0.00)
District Trend Effects	No	No	Yes
N	824	824	824

Notes: Observations are at the language-district level and are weighted by number of speakers in 1931. Regressions include controls for initial levels of urbanization, literacy, workforce share of population, and language share of population. Standard errors are corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 8: The Effects of Manufacturing Employment on Bilingualism: Weighted IV Estimates in Differences

	Δ Bilingual Share of Speakers	
	(1)	(2)
Δ Manufacturing Share of Emp.	0.372***	
	(0.09)	
Δ Manufacturing Share of Emp. × Minority Language	0.926***	0.592**
	(0.24)	(0.25)
Minority Language	-0.018	-0.036
	(0.02)	(0.02)
Constant	0.003	
	(0.01)	
District Trend Effects	No	Yes
R^2	0.609	0.656
N	824	824
Δ Manufacturing Share for Minority Language Speakers	1.298***	
	(0.28)	

Notes: Observations are at the language-district level and are weighted by number of speakers in 1931. Bilingual share of speakers is the ratio between the number of mother tongue speakers of language ℓ in district d who can speak a second language to the total number of mother tongue speakers of language ℓ in district d. Initial level controls include urbanization, literacy, workforce share of population, and language share of population. Standard errors are corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 9: The Effects of Agricultural Employment on Bilingualism: Weighted OLS Estimates in Differences

Panel A

	Δ Bilingual Share of Speakers	
	(1)	(2)
Δ Agriculture Share of Employment	-0.093***	
	(0.03)	
Δ Agriculture Share of Employment	-0.207***	-0.203**
× Minority Language	(0.08)	(0.08)
Minority Language	0.071	0.073
	(0.08)	(0.09)
Constant	0.058*	0.060***
	(0.03)	(0.00)
District Trend Effect	No	Yes
R^2	0.681	0.712
N	824	824

Panel B

	Δ Bilingual Share of Speakers	
	(1)	(2)
Δ Agricultural Laborers Share of Ag. Emp.	0.013	
	(0.02)	
Δ Agricultural Laborers Share of Ag. Emp.	0.167	0.201*
× Minority Language	(0.11)	(0.11)
Minority Language	-0.069*	-0.061
	(0.04)	(0.04)
Constant	-0.035***	0.059***
	(0.01)	(0.00)
District Trend Effects	No	Yes
R^2	0.668	0.703
N	824	824

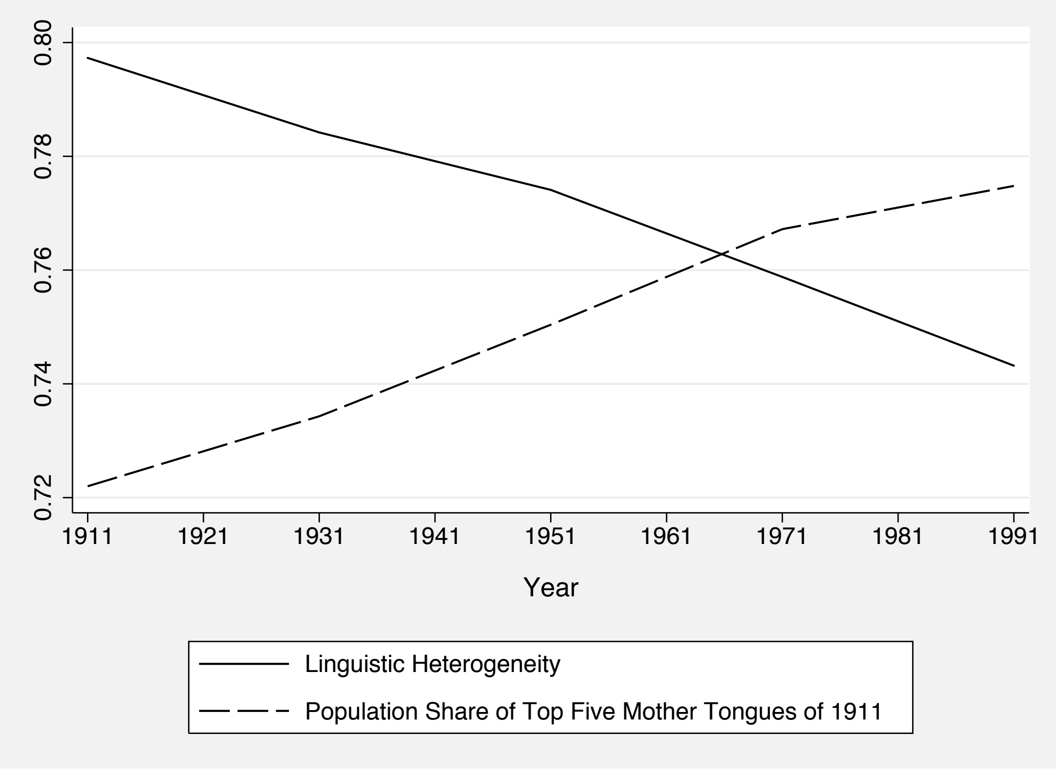
Notes: Observations are at the language-district level and are weighted by number of speakers in 1931. Bilingual share of speakers is the ratio between the number of mother tongue speakers of language ℓ in district d who can speak a second language to the total number of mother tongue speakers of language ℓ in district d. Initial level controls include urbanization, literacy, workforce share of population, and language share of population. Standard errors are corrected for clustering at the district level. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$.

Table 10: The Effects of Manufacturing Employment on Linguistic Heterogeneity

	Δ Linguistic Heterogeneity	
	(1)	(2)
Estimation	OLS	IV
Δ Manufacturing Share of Emp.	-1.559***	-1.503***
	(0.28)	(0.31)
Constant	0.090***	0.090***
	(0.03)	(0.03)
R^2	0.033	0.022
N	153	153

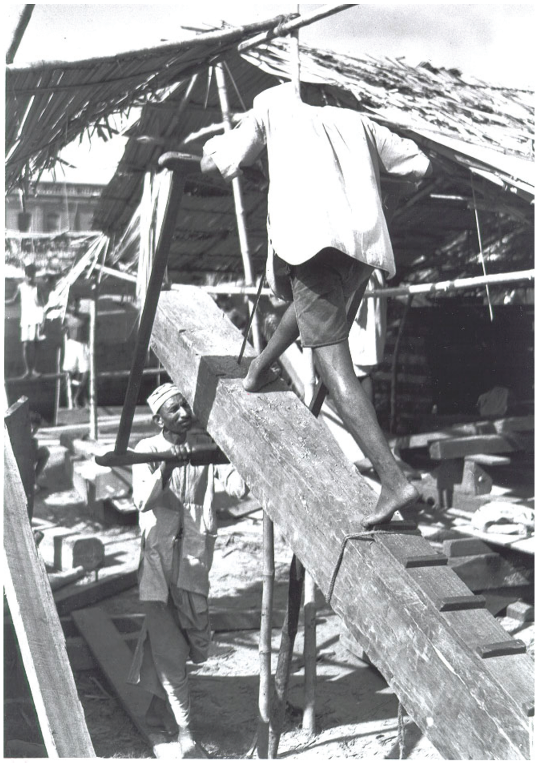
Notes: Each observation is weighted by the population in the district. Standard errors are robust to heteroskedasticity. Linguistic heterogeneity in district d is $h_d = 1 - \sum_\ell s_{\ell d}^2$, where $s_{\ell d}$ is the district population share of speakers whose mother tongue is ℓ. Regressions include controls for initial levels of urbanization, literacy, workforce share of population, and language share of population. Stars indicate statistical significance: * means $p < 0.10$, ** means $p < 0.05$, and *** means $p < 0.01$

Figure 1: Linguistic Consolidation in India, 1911–1991



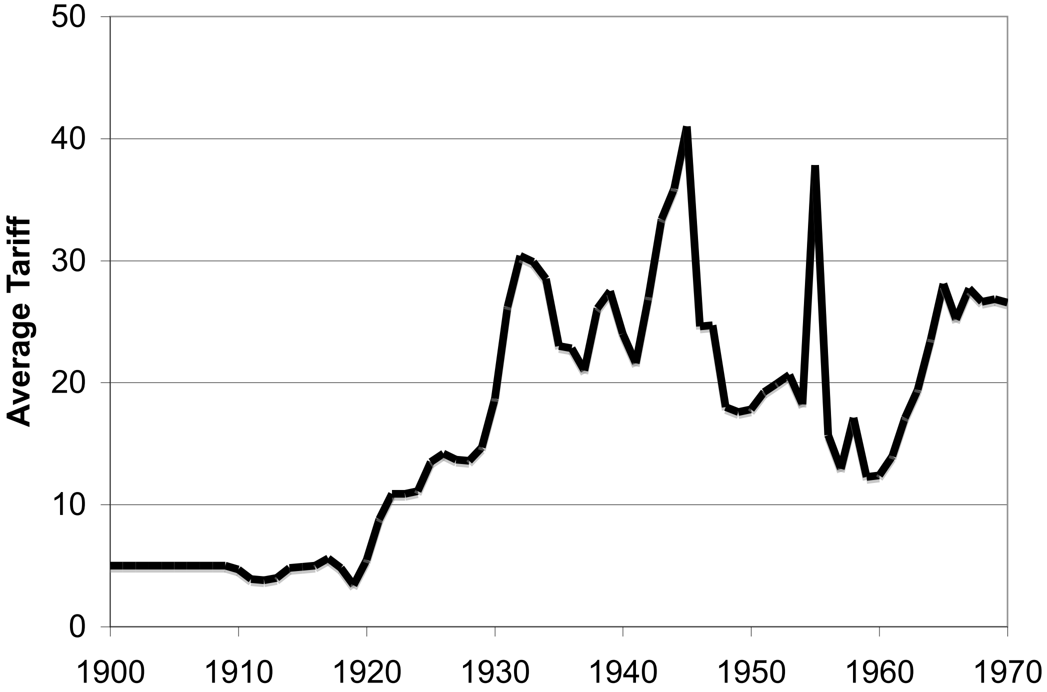
This graph shows two measures of linguistic consolidation for the post-Independence territory of India for the years 1911 to 1991. Linguistic heterogeneity is defined as $h = 1 - \sum_{\ell} s_{\ell}^{2}$, where s_{ℓ} is the population share whose mother tongue is ℓ. It measures the probability that two randomly selected individuals in the population will have different mother tongues. The population share speaking the top five mother tongues of 1911 measures the share of the population speaking Hindi, Marathi, Bengali, Tamil, and Telugu, which were the five largest mother tongues in 1911. Based on tables in the the 1961 and 1991 Census of India.

Figure 2: Sawyers at Work in a Boatyard, Calcutta, 1944



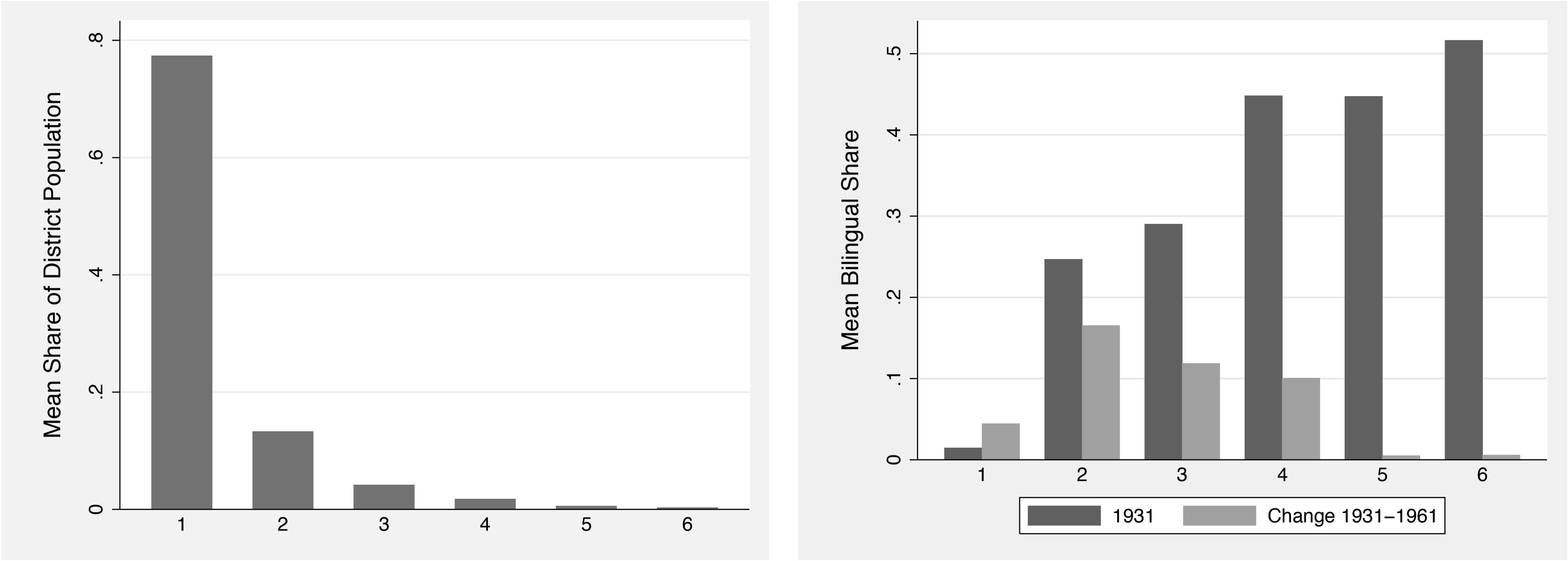
These two men are sawing a timber in a boatyard using the simple technology typical of the vast bulk of Indian industry in the mid-20th century. One man pushes down on the saw, and the other pushes up. (Hensley 1944).

Figure 3: Average Tariff in India, 1900–1970



Notes: The India tariff data in this chart comes from the database developed by Jeffrey Williamson and collaborators. Average tariffs are the ratio of the total import duties to the total value of imports in a given year.

Figure 4: District-Level Characteristics of Languages by their District-Level Ranks 1–6

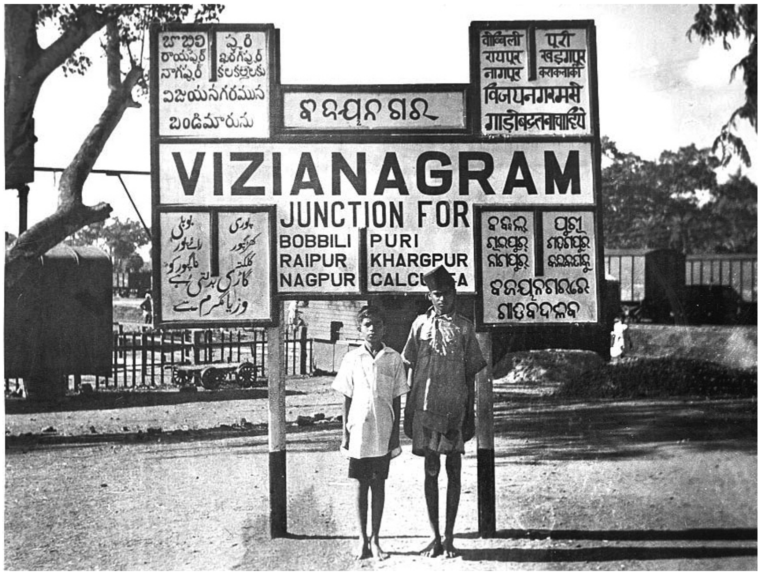


(a) Mean share of district population speaking language for each rank

(b) Mean district-level share of speakers of language bilingual in 1931 and change 1931–1961 for each rank

Notes: I ordered languages by their number of mother-tongue speakers within each district for 1931, assigning them a rank of 1 through 6. The right-hand graph shows the mean population share of languages by rank. The left hand graph shows the share of mother-tongue speakers who were bilingual in 1931 by the mother tongue's rank, and the change in bilingualism between 1931 and 1961. The left-hand graph is weighted by the number of speakers of the mother tongue of each rank in 1931. We can therefore interpret the graphs as showing information for the average person whose mother tongue has a given district-level rank.

Figure 5: Vizianagaram Station Sign in Five Languages, 1945



These two young porters stand in front of a station platform sign at Vizianagaram, a small town in the coastal region of Andhra Pradesh bordering Orissa. There are five languages on the sign. Beginning at the top left and reading across are Telugu, Oriya, and Hindi. English is in the center, and Urdu and Oriya are on the bottom panels. Interestingly, the town name is displayed on the top central panel in Oriya, the main language of Orissa (Hensley 1944).