

*Preliminary draft*

**LANGUAGE DIVERSITY AND ECONOMIC DEVELOPMENT**

**Paul De Grauwe**

**University of Leuven**

January 2006

I am grateful to Michel Beine, Hans Dewachter, Geert Dhaene, Marco Lyrio, Pablo Rovira Kaltwasser, Vivien Lewis, Agnieszka Markiewicz, Romain Houssa, Laura Rinaldi, Veerle Slootmaekers, Marijke Verpoorten, Blanca Zuluaga for many comments and suggestions. I am particularly indebted to Romain Houssa for helping me in collecting data.

## **1. Introduction**

The number of languages in the world is on a declining path. By some estimates, about 90 percent of the more than 6,000 languages spoken worldwide is endangered and is likely to disappear during this century (see Krauss (1995), Breton(1998)). Many observers have warned that as a result of massive language extinctions, the world's cultural diversity is threatened.

It has long been recognized that one of the fundamental causes of the decline in language diversity in the world is economic growth and development. The latter is seen as an unstoppable force that will lead to the disappearance of most living languages, much in the same way as it contributes to the reduction of the biodiversity of the world.

In this paper we analyze the link between economic development and language diversity in greater detail. One of the questions we will analyze is whether these dire predictions are warranted.

## **2. Economic development and language diversity : the theory**

The relation between economic development and language diversity is a complex one. One of the reasons is that the causality runs in two directions. There is a causal relation running from economic development to language diversity. This relation can be described as follows. Economic development is based on specialization and trade. Individuals who specialize and trade must develop common means of communication. This in turn leads to the use of a common language. Thus as countries move on the ladder of economic development and increase the network of trade both within and outside the country, a common language will impose itself and will be used by an increasing number of individuals. This then puts pressure on the local languages, and in the long run will push many of these into extinction. Thus in the long run economic development will lead to a decline in the number of languages and in language diversity.

The fact that language is characterized by network externalities reinforces this dynamics. With network externality we mean that the communication value (utility) of a language increases with the number of individuals who use that language

(Economides(1996)). Thus as one language increases in size, its communication value increases. As a result, the incentives to switch to its use by those who do not speak that language, increases. This process can, under certain conditions, lead to a situation where everybody uses the common language. The reverse side of this development is that the local languages tend to disappear and that language diversity declines.

There is also an inverse causality which in a way is quite obvious. The use of a common language intensifies trade because it facilitates communication. This link has been analyzed in great detail in the context of econometric analyses of international trade flows (see Eichengreen and Irwin(1998), Helliwell(1998), Mélitz(2005)). The results confirm the existence of a causal link from common language to trade, i.e. countries that speak the same language tend to trade more with each other than countries without a common language. To the extent that international trade promotes economic growth and development one can conclude that a common language also leads to more growth and development.

In this paper we intend to analyze the causal link running from economic development to language diversity. Such an analysis necessitates isolating the two different causal links. We will do this by introducing an appropriate instrumental variable.

Before starting the analysis it is useful to look at some broad data on the number and the distribution of languages in the world. In table 1 we show this information. It is striking to find that the (economically) least developed parts of the world (Africa and Oceania) are the habitat of about 50% of the spoken languages, while these regions represent only 12% of world population. Conversely, Europe that belongs to the most developed part of the world only has 3.5% of the world languages while it represents about 13% of world population.

**Table 1: Number of living languages in the world**

	number of living languages	percent	percent of world population
Africa	2092	30,3%	12,3%
Americas	1002	14,5%	14,4%
Asia	2269	32,8%	59,9%
Europe	239	3,5%	13,2%
Pacific	1310	19,0%	0,1%
Total	6912	100,0%	100,0%

*Source:* Gordon, Raymond G., Jr. (ed.), 2005.

### **3. Economic development and language diversity: econometric analysis**

We start by specifying an equation explaining the language diversity by the level of development, i.e.

$$L_i = a + b_1 Y_i + b_2 Z_i + \varepsilon_i \quad (1)$$

where  $L_i$  is the indicator of language diversity in country  $i$ ,  $Y_i$  is the level of development measured by per capita income in country  $i$ ,  $Z_i$  is a series of control variables (to be explained later), and  $\varepsilon_i$  is the error term.

We will use two alternative measures of language diversity. The first one is the number of languages spoken in a country. In this case one of the control variables is the size of the population. Clearly, the larger the population the more likely it is that many languages are spoken.

The previous measure of language diversity contains a bias in countries that are dominated by one language while at the same time having many small languages spoken by very few people. Typical examples are Latin American countries where Spanish or Portuguese dominates by far, while these countries also have many Indian languages being spoken in remote areas by small numbers of people. Brazil for example has about 200 spoken languages. However, Portuguese is the true common

language of Brazil and is spoken by 99% of the Brazilians. In order to correct for this bias, we will use a second indicator which in fact comes closest to what we wish to measure, i.e. an index of linguistic diversity (Greenberg diversity index). It is defined as

$$D = 1 - \sum_{k=1}^K (p_k)^2 \quad (2)$$

where  $p_k$  is the fraction of total population speaking language  $k$ ,  $K$  is the total number of languages in a country. Clearly when there is only one common language the index is 0. When the number of languages increases and the shares become more equal the index tends towards 1.

The index can be interpreted as representing the probability that any two persons of the same country selected at random would have different mother tongues. The highest possible value, 1, indicates total diversity (that is, no two people have the same mother tongue) while the lowest possible value, 0, indicates no diversity at all (that is, everyone has the same mother tongue).

Table 1 shows a few examples of these different indicators for a number of countries. It illustrates the contrast between Latin American countries and African and Asian countries with large numbers of languages. Typically, Latin American countries with large numbers of languages (e.g. Brazil and Mexico) have very low diversity indices while African and Asian countries where many languages are spoken have very high diversity indices.

**Table 2: Number of languages, mean number of speakers and diversity index**

	# languages	mean number of speakers	diversity index
Brazil	188	928.112	0,032
Mexico	291	346.457	0,135
Nigeria	505	262.941	0,87
India	415	2.526.846	0,93
Congo	218	236.606	0,948

Source: Gordon, Raymond G., Jr. (ed.), 2005.

We regress equation (1) using these two alternative indicators of language diversity. As control variables (variable  $Z_i$ ) we use population size and land size of country  $i$ .

As instrument variables we selected the World Bank's Human Development Index. We take the view that this variable has no direct influence on the language indicators, while it is strongly correlated with per capita income. It can therefore be used to isolate the causality going from income to language. In appendix we show the results of selecting other instruments, in particular those that relate to health and health care.

The sample consists of 167 countries. The data population, land size and per capita income were collected from the World Bank and relate to the year 2002. The language indicators are from Gordon, Raymond G., Jr. (ed.), 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL International. Online version: <http://www.ethnologue.com/>. All variables are expressed in logs (except for the index of diversity which is a number between 0 and 1).

We show the results of estimating equation (1) in table 3<sup>1</sup>. The results are quite encouraging: In the two specifications of the language indicator, per capita income has a highly significant effect on languages. More specifically, a 1 percent increase in per capita income leads to a decline in the number of languages by 0.23% and a reduction of the language diversity index by 0.11 points. This confirms that economic development significantly reduces linguistic diversity.

The control variables in most cases have the expected sign. Thus the size of population tends to increase the number of languages<sup>2</sup>. Land size increases the number of languages and increases linguistic diversity. The effect of population size on language diversity, however, is not statistically different from zero.

---

<sup>1</sup> In appendix we present the econometric results using OLS.

<sup>2</sup> Since we control for population the effect of per capita income can also be interpreted to mean that a 1% increase in per capita income increases the average size of a language, i.e. increases the mean number of people speaking a language.

**Table 3: Results of estimating equation (1)**

<b>Dependent Variable: Number of languages</b>				
Instrument list: C HDI LPOP LLAND				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.789	0.877	-2.03	0.0432
LPOP	0.233	0.072	3.22	0.0015
LYC	-0.226	0.059	-3.77	0.0002
LLAND	0.223	0.058	3.80	0.0002
R-squared	0.47	Mean dependent var		2.78
S.E. of regression	1.05	Durbin-Watson stat		2.00
<b>Dependent Variable: diversity index</b>				
Instrument list: C HDI LLAND LPOP				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.066	0.239	4.45	0.0000
LPOP	-0.013	0.019	-0.70	0.4824
LYC	-0.107	0.016	-6.59	0.0000
LLAND	0.036	0.016	2.26	0.0250
R-squared	0.18	Mean dependent var		0.45
S.E. of regression	0.28	Durbin-Watson stat		1.84

*Data sources:* World Bank (population, per capita income and land size)  
Gordon, Raymond G., Jr. (ed.), 2005.

The next step in the analysis consisted in checking the extent to which local conditions affect the results. Clearly, non-economic factors affect language developments on the different continents of the world. In order to capture these non-economic factors we introduced different dummy variables for each continent. The equation to be estimated now becomes

$$L_i = a + b_1 Y_i + b_2 Z_i + \sum_k b_k D_k + \varepsilon_i \quad (3)$$

where  $D_k$  is a dummy variables that obtains a value of 1 when the observation  $i$  belongs to continent  $k$ , and a zero otherwise. This variable captures the continent-specific effects on language diversity that are unrelated to per capita income, population and land size. We show the results of estimating equation (3) in table 4.

We observe from table 4 that the addition of continent dummies improves the fit considerably. Note that the constant measures the effect of the benchmark continent which is Africa. The other continent dummies then represent the effects of the other continents expressed as deviations from this benchmark. We observe that Europe and America have a lower linguistic diversity than Africa for reasons that are unrelated to

economic development, population size and land size. Asia and Oceania are not significantly different from Africa.

In general we also find that the addition of specific continent effects reduces the impact of per capita income (which remains significant, however). This suggests that without the continent dummies the income variable which is correlated with the continent dummies (e.g. per capita income in Africa, Asia and Oceania is very low) captures these non-economic continent-specific factors. It also suggests that there are powerful non-economic forces at work (e.g. differences in culture, geography) that lead to different degrees of linguistic diversity in different continents.

**Table 4: Results of estimating equation (3)**

<b>Dependent Variable: Number of languages</b>				
Instrument list: C HDI LPOP LLAND				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.262	0.932	-3.49	0.0006
LPOP	0.344	0.073	4.65	0.0000
LYC	-0.163	0.075	-2.15	0.0325
LLAND	0.166	0.056	2.93	0.0039
AMERICA	-0.004	0.251	-0.02	0.9852
ASIA	-0.173	0.222	-0.77	0.4366
EUROPE	-0.486	0.289	-1.68	0.0948
OCEANIA	1.558	0.418	3.72	0.0003
R-squared	0.55	Mean dependent var	2.78	
S.E. of regression	0.98	Durbin-Watson stat	1.98	
<b>Dependent Variable: diversity index</b>				
Instrument list: C HDI LLAND LPOP				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.909	0.246	3.68	0.0003
LPOP	-0.018	0.019	-0.94	0.3453
LYC	-0.067	0.020	-3.37	0.0009
LLAND	0.038	0.015	2.54	0.0119
AMERICA	-0.301	0.066	-4.54	0.0000
ASIA	-0.029	0.058	-0.51	0.6112
EUROPE	-0.139	0.076	-1.82	0.0704
OCEANIA	0.003	0.110	0.025	0.9802
R-squared	0.35	Mean dependent var	0.45	
S.E. of regression	0.26	Durbin-Watson stat	1.97	

*Data sources:* World Bank (population, per capita income and land size)  
Gordon, Raymond G., Jr. (ed.), 2005.

#### 4. A counter-factual exercise

The previous empirical estimates (equation (3), table 4) allow us to perform a counter-factual exercise. This consists in analyzing what number of languages the model predicts for different levels of per capita income. We structured the counter-factual as follows. We computed how many languages each continent would have under two counter-factual levels of per capita income. In the first counter-factual we assumed that each continent has the average European per capita income. Under the second counter-factual we assumed that each continent has the per capita income of France. We show the results in table 5. The first row is the observed number of languages. With observed is meant here the number of languages as predicted by the model given the observed per capita incomes in the different continents. The next rows show the number of languages as predicted by the model (and the percentage decline relative to the observed ones) using the European and French per capita incomes, respectively.

The results show that if the other continents “graduate” to the European and French levels of income the total number of languages will decline significantly. The largest part of the decline is in Africa, the continent with the lowest per capita income. If Africa had the French per capita income, the number of languages would decline by more than 40%.

These are significant effects. Yet they fall short from the alarmist predictions quoted in the introduction and suggesting that before the end of the century 90% of the world’s languages may disappear. If these predictions turn out to be correct, other than economic forces will have to play a role.

**Table 5: Number of languages assuming different per capita income levels**

	Africa	America	Asia	Europe	Oceania	Total
number of languages						
observed	1704	1221	2028	563	900	6416
assuming European income <i>percent decline</i>	1096 35,6%	1003 17,8%	1600 21,1%	563 0,0%	792 12,0%	5055 21,2%
assuming French income <i>percent decline</i>	1001 41,3%	916 25,0%	1460 28,0%	514 8,7%	723 19,7%	4614 28,1%

#### **4. Conclusion**

It has long been recognized that economic development puts pressure on local languages. The mechanism producing this pressure has to do with the fact that economic development is based on specialization and trade, which in turn requires a common language as a means of communication. Thus as economic development proceeds more and more people take on a common language thereby reducing the importance of local languages and contributing to their extinction. According to some linguists, this process may eliminate up to 90% of the world's living languages before the end of this century.

The purpose of this paper was to test the hypothesis that economic development tends to reduce linguistic diversity. The difficulty in testing this hypothesis is that the causal relation between economic development and linguistic diversity runs in two directions. While economic development reduces linguistic diversity, it is also true that the existence of a common language enhances economic development. The latter causal relation arises because the inhabitants of countries with little linguistic diversity can communicate better, thereby increasing the scope for trade and specialization.

In order to separate out these two causal relations we used instrumental variables. We selected the Human Development Index of the World Bank for that purpose. Using a sample of more than 160 countries, and controlling for a number of other factors, we found that economic development (measured by per capita income) has a significant and powerful effect on linguistic diversity. We then simulated two counter-factual scenarios in which we asked the question what the number of languages would be if all the countries in the world had the same per capita income of the average European country and of France, respectively. We found that in these counter-factual scenarios the number of languages in the world would be 20 to 30% lower. The effect is most pronounced for Africa where the number of languages would decline by 35-40%.

Although economic development is important in affecting linguistic diversity, there are other non-economic forces at work. These tend to dampen the effect of economic development on linguistic diversity. This finding led us to conclude that the prediction that the number of living languages in the world would decline by 90% before the end of the century is probably exaggerated.

The results reported here should be considered as preliminary. An important limitation of the analysis is that we used cross-section data. These have no time dimension. It would be interesting to add a time dimension and to estimate panel data models. The problem with this is that there is very little information on the evolution of language diversity over time. This contrasts with the information we have about the evolution in per capita income. It will therefore remain difficult to introduce a time dimension in linguistic diversity without losing information about many countries.

## References

- Breton, A., (1998), An economic analysis of language, in Breton, A., ed., *Economic Approaches to Language and Bilingualism*, Canadian Heritage, Ottawa, 1-33.
- Economides, N., (1996), The economics of networks, *International Journal of Industrial Organization*, 14, 673-99.
- Eichengreen, B., and Irwin, D., (1998), The Role of History in Bilateral Trade Flows, in Frankel, J., ed., *The Regionalisation of the World Economy*, Chicago University Press, 33-57.
- Gordon, R., Jr. (ed.), 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL International. Online version: <http://www.ethnologue.com/>
- Helliwell, J., (1998), *How Much Do National Borders Matter?*, Brookings Institution, Washington, D.C.
- Krauss, M., (1995), *Endangered Languages: Current Issues and Future Prospects*, Keynote Address, Dartmouth college, Hanover, NH, 3 February
- Mélitz, J., (2005), *Language and Foreign Trade*, CEPR Working Paper, no. 3590, London..
- World Bank, (2005), *World Development Report*, Washington, D.C.

## Appendix 1

This appendix presents results using health indicators as instrumental variables. We use both health spending (as a percent of GDP) and infant mortality rate. The source of these data is the World Bank. The results are shown in table A1

**Table A1: Estimating equation (1) using health indicators as instrumental variables**

<b>Dependent Variable: Number of languages</b>				
Instrument list: C LHEALTH LINFANT LPOP LLAND				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.768	0.852	-2.07	0.0397
LPOP	0.198	0.071	2.77	0.0062
LYC	-0.187	0.058	-3.19	0.0017
LLAND	0.242	0.059	4.09	0.0001
R-squared	0.47	Mean dependent var		
S.E. of regression	1.051	Durbin-Watson stat		1.94
<b>Dependent Variable: diversity index</b>				
Instrument list: C LHEALTH LINFANT LPOP LLAND				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.899	0.229	3.92	0.0001
LPOP	-0.013	0.019	-0.72	0.4677
LYC	-0.091	0.015	-5.79	0.0000
LLAND	0.040	0.016	2.50	0.0131
R-squared	0.21	Mean dependent var		0.45
S.E. of regression	0.28	Durbin-Watson stat		1.89

We also performed regression using the same health indicators as instrumental variables together with Continent dummies. The results are shown in table A2. On the whole the results are very similar to those reported in the main text. The coefficients of per capita income, however, turn out to be smaller.

**Table A2: Estimating equation (3) using health indicators as instrumental variables**

<b>Dependent Variable: Number of languages</b>				
Instrument list: C LINFANT LLAND LPOP AMERICA ASIA EUROPE OCEANIA				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.626	0.924	-3.92	0.0001
LPOP	0.323	0.074	4.35	0.0000
LYC	-0.104	0.074	-1.40	0.1630
LLAND	0.191	0.057	3.34	0.0010
AMERICA	-0.077	0.253	-0.30	0.7590
ASIA	-0.250	0.225	-1.11	0.2685
EUROPE	-0.608	0.288	-2.10	0.0366
OCEANIA	1.316	0.387	3.40	0.0009
R-squared	0.542580	Mean dependent var	2.759271	
S.E. of regression	0.981102	Durbin-Watson stat	1.986773	
<b>Dependent Variable: diversity index</b>				
Instrument list: C LINFANT LLAND LPOP AMERICA ASIA EUROPE OCEANIA				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.746	0.240	3.10	0.0023
LPOP	-0.020	0.019	-1.08	0.2798
LYC	-0.045	0.019	-2.34	0.0203
LLAND	0.043	0.014	2.93	0.0038
AMERICA	-0.341	0.065	-5.20	0.0000
ASIA	-0.064	0.058	-1.10	0.2718
EUROPE	-0.200	0.074	-2.67	0.0082
OCEANIA	-0.118	0.100	-1.18	0.2392
R-squared	0.37	Mean dependent var	0.45	
S.E. of regression	0.26	Durbin-Watson stat	1.97	

## Appendix 2: OLS results

<b>Dependent Variable: Number of languages</b>				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.040	0.808	-2.52	0.0126
Population	0.193	0.070	2.74	0.0068
Per Capita Income	-0.151	0.052	-2.90	0.0042
Land size	0.249	0.058	4.29	0.0000
R-squared	0.47	Mean dependent var		2.75
S.E. of regression	1.047	Prob(F-statistic)		0.00
Durbin-Watson stat	1.98			
<b>Dependent Variable: diversity index</b>				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.753	0.216	3.48	0.0006
Population	-0.016	0.018	-0.85	0.3936
Per Capita Income	-0.073	0.013	-5.29	0.0000
Land size	0.043	0.015	2.82	0.0053
R-squared	0.23	Mean dependent var		0.45
S.E. of regression	0.28	Prob(F-statistic)		0.000
Durbin-Watson stat	1.92			