The role of language in shaping international migration: Evidence from OECD countries 1980-2010

(Results in current preliminary version are based on data 1985-2006; currently we are extending the dataset as noted in the abstract)

Alicia Adsera (Princeton University and IZA) and Mariola Pytlikova (Aarhus University, CCP and CIM, Denmark)

1. INTRODUCTION

Previous literature has shown that both fluency in the language of the destination country or the ability to learn it quickly as well as whether that language is a widely spoken language in the world play a key role in the transfer of existing human capital to a foreign country and generally boost the immigrant's success at the destination country's labor market, see e.g. Kossoudji (1988), Bleakley and Chin (2004); Chiswick and Miller (2002, 2007, 2010), Dustmann (1994), Dustman and van Soest (2001 and 2002), Dustmann and Fabbri (2003), Adsera and Chiswick (2007) and Toomet (2011). Thus linguistic skills and linguistic proximity seem to be very important in accounting for migrants' well-being. This suggests that the ability to learn and speak a foreign language quickly might be an important factor in the potential migrants' decision making. However, previous evidence on the determinants of migration typically included only a simple dummy for sharing a common language[1].

The main contribution of this paper is to investigate in depth the role of language in shaping international migration by using a wide range of linguistic indicators. First, we examine the relevance of linguistic proximity between origin and destination countries in the decision to migrate and to this aim we construct a refined indicator of the linguistic proximity between two countries based on the linguistic family of both the official and any other local language in each country. In addition, we employ the linguistic proximity measure proposed by Dyen et al. (1992), a group of linguists who built an index of distance between Indo-European languages based on the similarity

[1] A few studies have also employed some more sophisticated linguistic measures. For instance Adsera and Chiswick(2007) use dummies for broad language families (i.e. romance group) in their earnings equation and they find an earnings premium for immigrant men to EU countries arriving from a country whose official language belongs to the same linguistic family of the language in the destination country. Further there are two studies that use more complex linguistic proximity measures in studying migration determinants. Belot and Hatton (2011) use the number of nodes between one language and another on the linguistic tree to construct a linguistic proximity measure. Finally, a recent paper by Belot and Ederveen (2010) employs the linguistic proximity index proposed by Dyen et al. (1992). The authors show that cultural barriers explain patterns of migration flows between developed countries better than traditional economic variables. In our paper, we use the Dyen index as a part of robustness analyses.

between samples of words from each language. To separate the relevance of language proximity on its own from other sources of cultural proximity we also include information on genetic distance between destination and source country's populations in the models. Second, we investigate the hypothesis that potential migrants prefer to choose a destination with a "widely spoken" language, such as English, as its local language. The rationale behind this is the following: knowledge of particular foreign languages enhances the chances of success of a potential immigrant at the foreign labour market and lowers his/her costs of migration. Further, foreign language proficiency might be considered an important part of a worker's human capital in the labour market of the source country (European Commission, 2002 and Toomet, 2011). Thus, learning/practicing/improving a "widely spoken" language in the destination country serves as a pull factor especially for temporary migrants. Third, we investigate the role of the richness and variety of the linguistic environment at destination and origin in the migration process. Numerous neuroscience and biology studies have argued that a multilingual environment may shape brains of children differently and increase capacity to absorb further more languages (Kovacs and Mehler, 2009). If this is the case we should expect, ceteris paribus, lower costs of migration for people from multi-lingual countries, and consequently larger emigration fluxes from those countries. Regarding the effect of linguistic diversity and polarization at the destination country on migration, there might be two forces pulling the effect into different direction: a linguistically polarized society may increase the costs of adaptation, but a diverse society might have in place more flexible policies that adapt to the needs of different constituencies (e.g. education, integration programs). We also add to the existing literature on determinants of migration by analyzing a rich international migration dataset, which allows us to analyze migration from a multi-country perspective. In this paper, we analyze determinants of gross migration flows from 130 countries to 27 OECD countries annually for the period 1985-2006.

We find that emigration rates are higher among countries whose languages are more similar. The result is robust to the inclusion of genetic distance, which suggests language itself affects migration costs beyond any ease derived from moving to a destination where people may look or be culturally more similar to the migrant. We conduct the analysis by looking at both the proximity between the most used languages in each country as well as to the maximum proximity between any of the official languages (if multiple) in both countries. Among countries with Indo-European languages this result is highly robust to the use of an alternative continuous measure of proximity developed by linguists. When splitting the sample for English and non-English speaking destinations, linguistic proximity matters significantly for the latter group. The average migrant likely has some English proficiency, even before the move, that may temper the relevance of the linguistic proximity when studying flows to English speaking destinations. Finally, destinations that are linguistically more diverse and polarized attract fewer migrants than those with a single language; whereas more linguistic polarization at origin seems to act as a push factor.

The rest of the paper is organized as follows: Section 2 surveys earlier research in the area and presents the theoretical framework of the paper. Section 3 shortly presents a model on international migration on which we base our empirical analysis. Section 4 describes the empirical model as well as the database on migration flows and stocks collected for this study and the independent variables included in the analyses. Results from the econometric estimates are given in Section 5. Finally, Section 6 offers some concluding remarks.

2. THEORY AND PREVIOUS RESEARCH ON MIGRATION DETERMINANTS

2.1 Migration Determinants and Linguistic Proximity

The determinants and consequences of migratory movements have long been discussed in the economic literature. The first contributions can be found in neoclassical economics, which stress differentials in wages as a primary determinant of migration (Hicks, 1932). The “human capital investment” theoretical framework (Sjaastad, 1962) adds migration costs to the model of migration, so that a person decides to move to another country only if the discounted expected future benefit of moving is higher than the cost of migration. The “human capital investment” model has been further adjusted by including the probability of being employed in each location; see Harris and Todaro (1970). In aggregate terms, the differentials in wages and probability of unemployment are typically proxied by GDP per capita levels and unemployment rates in destination and source countries, respectively. The effect of GDP per capita in the source country on migration flows may be mixed since poverty constrains the ability to cover costs of migration. It has been shown in previous studies, e.g. Chicquar and Hanson (2005), Hatton and Williamson (2005), Clark, Hatton and Williamson (2007), Pedersen et al. (2008) and Vogler and Rotte (2000), that source country’s GDP per capita has an inverted U-shape effect on migration[2].

[2] At income levels beyond dire poverty, migration increases, but after GDP reaches a certain level, migration may again decrease because the economic incentives to migrate to other countries decline.

In addition to the economic determinants, Borjas (1999) argues that generous social security payment structures may play a role in migrants' decision making. The idea behind this is that potential emigrants must take into account the probability of being unemployed in the destination country. The damaging consequences of unemployment may be reduced with the existence of generous welfare benefits in the destination country. Such welfare transfers constitute basically a substitute for earnings during the period devoted to searching for a job. However, empirical studies are not conclusive in this respect; see e.g. Zavodny (1997), Pedersen et al. (2008), among others. Besides, immigration policies and changes in these policies over time strongly contribute to shape migration flows as their impact among individuals from different source countries for each potential receiving country may differ (Clark et al. 2007; Mayda, 2010; Ortega and Peri 2009).

The costs of migration are also an important part of migrants' decision making. They include not only the immediate out-of-pocket expenses, but also psychological costs connected to moving to a foreign country and leaving behind family, friends and a familiar environment. Costs typically increase with the physical distance between two countries. However, changes and improvements in communication technologies and declining transportation prices may have reduced the relevance of physical "distance" during the latest decades. Further, network effects may also counteract the deterrent effect of "distance". Through "networks" potential migrants receive information about the immigration country - about the likelihood of getting a job, economic and social systems, immigration policy, people and culture. This facilitates the move and the adaptation of new immigrants into the new environment (Massey et al. 1993; Munshi, 2003). Network effects may also help to explain the persistence of migration flows; see e.g. Bauer et al. (2005, 2007), Heitmueller (2006) and Clark et al. (2007). Empirical evidence has shown that migrant networks have a significant impact on sequential migration, see e.g. Pedersen et al. (2008), who also show that networks are more important to people coming from low-income developing countries compared to migrants originating from high-income countries. The latter is also supported by McKenzie and Rapoport (2010) and Beine et al. (2011) who find that diasporas explain a majority of the variability and selection in migration flows.

In addition, the linguistic and cultural distance between source and destination country is as well important. The more "foreign" or distant the new culture and the larger the language barriers are, the higher are the migration costs for an individual and the less likely it is that the individual decides to migrate, holding all other factors constant (Pedersen et al., 2008). A recent study by Belot and Ederveen (2010) shows that cultural barriers, as measured by a diverse set of cultural, religious and linguistic distance indexes based on Hofstede (1991), Baker and Inglehart (1991) and Dyen et al. (1992), explain patterns of migration flows between developed countries better than traditional economic variables.

In particular, the ability to speak a foreign language is an important input in the decision process of a potential migrant. Fluency in the language of the country of destination or in widely spoken languages plays a key role in the ability to transfer human capital from the source to a foreign country and generally helps the immigrant to succeed at the destination’s labor market, see e.g. Kossoudji (1988), Bleakley and Chin (2004); Chiswick and Miller (2002, 2007), Dustmann (1994), Dustman and van Soest (2001 and 2002) and Dustmann and Fabbri, (2003). By exploiting differences between young and old arrivers from non-English speaking source countries on their adult English proficiency, Bleakley and Chin (2004 and 2010) find that linguistic competence is a key variable to explain immigrant’s disparities in terms of educational attainment, earnings and social outcomes. Adsera and Chiswick (2007) found that there is around a 9 % earnings premium for immigrant men to EU countries arriving from a country whose official language belongs to the same linguistic family of the language in the destination country. Besides, a “widely-spoken” native language in the destination country can constitute a pull-factor in international migration. Two different forces may lie behind that migration pattern. First, as some “widely spoken” languages are often taught as second languages at schools in many source countries, the immigrants are more likely to migrate to destinations, where those languages are spoken. Second, foreign language proficiency is considered an important part of human capital in the labor market of the source country, see e.g. European Commission (2002) on language proficiency as an essential skill for finding a job in home countries. A recent article by Toomet (2011) finds that knowledge of English is associated with a 15% wage premium on the Estonian labor market. Thus, learning/practising/improving the skills of “widely spoken” languages in the destination countries serve as a pull factor especially for temporary migrants.

2.2 Linguistic Diversity and Polarization

Additionally the richness and variety of the linguistic environment where an individual is brought up may enhance his/her future ability to adapt to a new milieu. Numerous neuroscience and biology studies have argued that a multilingual environment may shape brains of children differently and increase their capacity to absorb better a larger number of languages (Kovacs and Mehler, 2009). If this is the case we should expect that, ceteris paribus, individuals from multi-lingual countries would have an easier time absorbing a new linguistic register in their destination country. In that regard the migration costs of those individuals would be smaller than otherwise and we would expect larger immigration fluxes (and better outcomes, something beyond the scope of this paper) from those source countries, other things being constant.

At the same time an increase in the diversity of languages at origin may also be a proxy for ethnic or political fractionalization that can by itself act a push factor for migration out of the country. Some literature argues that ethnic fractionalization has been conducive to more internal conflicts or civil wars (though the literature is still contentious over this issue, e.g. Fearon 2003) and may lead to more inefficient allocation of resources that deter growth. In that regard, how large the different linguistic groups within a country are and how wide their linguistic distances are could be related to whether political tension may be associated or not with linguistic diversity. A set of existing measures of polarization, developed from the initial work of Esteban and Ray (1994) and Duclos et al. (2004), are able to capture this dimension of diversity. Esteban and Ray (1994, 2006) and Montalvo & Reyal-Querol (2005) have shown polarization to be relevant, beyond traditional measures of inequality or diversity (e.g. income, ethnic groups...), to understand political demands and civil strives, among other things. Similarly Desmet et al. (2009) and Desmet et al. (2011) measure ethno-linguistic diversity and offer new results linking such diversity with a range of political economy outcomes -- civil conflict, redistribution, economic growth and the provision of public goods. In the empirical analysis of this paper we use measures of both diversity and polarization developed by Desmet et al. (2009) that take into account linguistic distances across the different groups in a society to understand whether some of those forces may be at play in migration decisions.

We posit that larger linguistic polarization might act as a push factor migration, even after controlling for income levels and the degree of political freedom, since it may be correlated with lower inter-group trust among other things.

Similarly, the degree of diversity and/or polarization of languages at the destination country may make it more or less attractive to the potential migrant. A largely polarized society may increase the costs of adaptation, even after linguistic distance between the language of the migrant and the dominant (or the closest) language of the destination country is taken into account. Further, linguistic diversity should not pose the same problems if the linguistic distance of the different linguistic groups in a country is not that large than if it is sizable. Finally, a diverse society might have in place more flexible policies that adapt to the needs of different constituencies (e.g., education immersion in different languages across different areas in the country to facilitate adaptation of newcomers such as those in place in Quebec).

Although the role of language and linguistic proximity seem to be very important, previous evidence on the determinants of migration hardly ever went beyond the inclusion of a simple dummy for sharing a common language. This paper contributes to the literature by exploring the different dimensions of the link.

3. A MODEL OF INTERNATIONAL MIGRATION

The standard neoclassical theory assumes that potential migrants have utility-maximizing behaviour, that they compare alternative potential destination countries and choose the country, which provides the best opportunities, all else being equal. Immigrants' decision to choose a specific destination country depends on many factors, which relate to the characteristics of the individual, the individual's country of origin and all potential countries of destination. Following Zavodny (1997), Karemera et al. (2000) and Pedersen et al (2008), we consider individual *k's* expected utility in country *j* at time *t* given that the individual lived in the country *i* at time *t-1*

$$U_{ijkt} = U(X_{ikt-1}, X_{jkt-1}, S_{ijkt-1}, D_{ij}) \tag{1}$$

where X_{ikt-1} and X_{jkt-1} are vectors of push and pull factors that vary across time and affect individual *k's* choice. The vector D_{ij} includes time-independent fixed-out-of-pocket and psychological/social costs of moving from country *i* to country *j*. The vector S_{ijkt-1} includes information on the individual's available network connections that affect his utility of living in country *j* at time *t,* given that the individual lived in country *i* at time *t-1*. For example, an individual may want to move to a country where his friends, family members or country fellows are. We assume the utility of an individual has a linear form:

$$U_{ijkt} = \alpha_1 S_{ijkt-1} + \alpha_2 D_{ij} + \alpha_3 X_{ikt-1} + \alpha_4 X_{jkt-1} + \varepsilon_{ijkt} \tag{2}$$

where ε_{ijkt} represents an idiosyncratic error term and $\alpha_1, \alpha_2, \alpha_3$ and α_4 are vectors of parameters of interest to be estimated, *i* denotes source country and *j* denotes destination country, (*i* = 1,…,130, and *j* = 1,….,27); *t* is time period (*t* = 1,…,22). A potential immigrant maximizing his utility chooses the country with the highest utility at time *t* conditional on living in country *i* at time *t-1*. Thus, we can write the conditional probability of individual *k* choosing country *j* from 27 possible choices as:

$$\Pr(j_{kt} / i_{kt-1}) = \Pr\left[U_{ijkt} = \max(U_{ki1t}, U_{ki2t}, ..., U_{ki27t})\right] \qquad (3)$$

Model (3) might be used to estimate the determinants of the individual's locational choice. However, as we use macro data, we aggregate up to population level by summing over *k* individuals. The number of individuals migrating to country *j,* whose utility is maximized in that country, is given by:

$$M_{ijt} = \sum_k \Pr\left[U_{ijkt} = \max(U_{ki1t}, U_{ki2t}, ..., U_{ki27t})\right] \qquad (4)$$

where M_{ijt} is the number of immigrants moving to country *j* from country *i* at time *t*. We assume a linear form of the variables that influence the locational choice of immigrants. Hence we have:

$$M_{ijt} = \beta_1 S_{ijt-1} + \beta_2 D_{ij} + \beta_3 X_{it-1} + \beta_4 X_{jt-1} + \mu_{ijt}, \qquad (5)$$

where μ_{ijt} is an error term assumed to be *iid* with zero mean and constant variance.

Next section presents the dataset employed in the analysis as well as the particular empirical specification used.

4 EMPIRICAL MODEL SPECIFICATION

4.1 Data

The analysis is based on data on immigration flows and stocks of foreigners in 27 OECD destination countries from 130 source countries for the years 1985–2006. The original OECD migration dataset by Pedersen, Pytlikova and Smith (2008) covered 22 OECD destination and 129 source countries over the period of years 1989-2000 (see Pedersen, Pytlikova and Smith (2008) for a detailed description of the dataset). For the purpose of this paper we additionally included

Slovenia as country of origin and collected data from 5 other OECD countries as additional destinations – Czech and Slovak Republics, Hungary, Poland and Ireland. Further, we extended the existing time period to include the years 1985-1989 and 2001-2006. The dataset has been collected by writing to selected national statistical offices of the 27 OECD countries to request detailed information on immigration flows and foreign population stocks by source country in their respective country.

Besides the information on flows and stocks of migrants, the dataset contains a number of other time-series variables, which may help to explain the migration flows between countries. These variables were collected from various sources (e.g. OECD, the World Bank and others). The Appendix contains definitions, sources of the variables and summary statistics. Although our data set presents substantial progress over that used in past research, there are still some problems related to its nature. First of all, the data set is unbalanced, with some missing information on migration flows and stocks, but also on some explanatory variables mostly for countries of origin. Another important problem is that, different countries use different definitions of an "immigrant" and different sources for their migration statistics.[3] In definitions of immigration flows some countries like Australia, Canada, Ireland, the Netherlands, Poland and the United States define an "immigrant" by country of birth. Other countries like New Zealand, The Slovak Republic, and Spain use definition by country of origin, while the rest of countries define an immigrant by citizenship. For immigration stock, the definition of immigrant population differs among countries as well, but for the majority of destinations we use the definition by country of birth.[4] For a more comprehensive description of the dataset, see Pedersen, Pytlikova and Smith (2008).

4.2. Empirical model

Departing from equation (5), we normalize the immigration flows by population size of the source country and we use the emigration rate, m_{ijt} instead of migration flow in absolute numbers as the

[3] For example, Belgium, Germany, Luxembourg, the Netherlands, Switzerland and the Nordic countries use data based on population registers; the majority of Southern and Eastern European countries use data based on the number of residence permits extended; Australia, Canada, New Zealand and Poland use data from censuses; some countries like Greece, the United Kingdom and the United States use labour force surveys and others have information based on social security systems or other sources.

[4] The majority of countries, in particular Australia, Austria, Canada, Denmark, Finland, France, Iceland, Ireland, the Netherlands, New Zealand, Norway, Poland, the Slovak Republic, Spain, Sweden, the United Kingdom and the United States define immigrant population by country of origin or country of birth. A few countries like Belgium, Czech Republic, Germany, Greece, Hungary, Italy, Japan, Luxembourg, Portugal and Switzerland define immigrant population by citizenship..

dependent variable. Further, we also normalize the lagged stock of immigrants, our proxy for "networks", by dividing the stock by the population of the source country i, s_{ijt-1} .

A model with emigration rate on the left hand side and a number of explanatory pull and push factors, as well as distance variables on the right hand side constitutes our basic gravity model of immigration. All variables used in the estimations except dummy variables are expressed in logarithms and the estimated coefficients represent impact elasticities. Further, in order to account for the information available to the potential migrant at the time the decision whether to move or not was made the relative differences in economic development and employment between origin and destination countries should be lagged. More importantly, there might be a problem of reverse causality if migration flows impact earnings and employment.[5] Lagging the economic explanatory variables and treating them as predetermined is one way to reduce the risks of reverse causality in the model[6]..

Thus, the model to be estimated is:

$$\ln m_{ijt} = \beta_1 \ln s_{ijt-1} + \beta_2 X_{it-1} + \beta_3 X_{jt-1} + \beta_4 D_{ij} + \beta_5 L_{ij} + c_j + c_i + \mu_{ijt} \qquad (6)$$

The explanatory variables included in X_{it-1} and X_{jt-1} cover a number of push and pull factors such as economic development measured by GDP per capita in destination and source countries (to capture the relative income opportunities in the two countries), employment prospects in the sending and receiving countries, measured by unemployment rates, and relative size of populations in destination and source countries. As an additional pull factor we include information on the extent of welfare provisions in the country of destination measured by public social expenditure as percentage of GDP. Political pressure in the source country may also influence migration. Therefore, we include a couple of indices from *Freedom House* which intend to measure the degree of freedom in first, political rights and second, civil liberties in each country. Each variable takes on values from one to seven, with one representing the highest degree of freedom and seven the lowest. Violated political rights and civil liberties are expected to increase migration outflows in a given country. All pull and push variables are in logs.

[5] There is another huge stream of literature that focuses on the effect of immigration on the labour market, see e.g. Chiswick (1996), Filer (1992), Hunt (1992), Friedberg and Hunt (1995), Chiswick and Hatton (2002), Borjas (2003), Card, (2005), Ottaviano and Peri (2005 and 2010), Hanson (2009), D'Amuri et al. (2010) and Peri (2010).

[6] With regard to the migrants' network, the variable is endogenous since the stock is just a function of previous stock plus migration flows minus out-migration, and therefore, we also lag the stock of migrants and assume that the lagged stock is predetermined with respect to the migration flows.

Matrix D_{ij} contains distance variables reflecting costs of moving to a foreign country. First, we include a dummy variable to proxy for cultural similarity denoted *Neighbour Country* which takes the value of 1 if the two countries are neighbours, and 0 otherwise. The variable *Colony* is a dummy variable with the value of 1 for countries ever in colonial relationship, and 0 otherwise. This variable is included because past colonial ties might have some influence on the cultural distance between countries, increase the information available and general knowledge about the potential destination country in the source country and thus lower migration costs and encourage migration flows between these countries. In order to control for the direct costs (transportation costs) of migration we use the measure of the *Log Distance in Kilometres* between the capital areas in the sending and receiving countries.

In most models we include a full set of destination and source fixed effects, $c_{j,}$ and $c_{i,}$ in order to capture unobserved factors influencing immigration flows such as differences in national immigration policy, or climate.

The linguistic variables of interest in this paper are covered in matrix *L*. First, we include a variable *Linguistic Distance,* an index ranging from 0 to 1 depending on how many linguistic families the languages of both the destination and the source country belong to. We constructed the index in the following way: first we defined 4 weights, equal to 0.1 if two languages are only related at the most aggregated linguistic tree level, e.g. Indo-European versus Uralic (Finnish, Estonian, Hungarian); 0.15 if two languages belong to the same second- linguistic tree level, e.g. Germanic versus Slavic languages; 0.20 if two languages belong to the same third linguistic tree level, e.g. Germanic West vs. Germanic North languages; and 0.25 if both languages belong to the same fourth (highest) level of linguistic tree family, e.g. Scandinavian West (Icelandic) vs. Scandinavian East (Danish, Norwegian and Swedish), German vs. English, or ItaloWest (Italian, French, Spanish, Catalan and Portuguese) vs. RomanceEast (Romanian). Next, we create the linguistic proximity index as a sum of those weights above[7], and we set the index equal to 0 if two languages do not belong to any common language family, and equal to 1 for a common language in two countries. Thus the linguistic proximity index equals 0.1 if two languages are only related at the most aggregated linguistic tree level, e.g. Indo-European languages ; 0.25 if two languages belong to the same first and second- linguistic tree level, e.g. Germanic languages; 0.45 if two languages share the same first up to third linguistic tree level, e.g. Germanic North languages; and 0.7 if both languages share

[7] Thus the index is equal to: 0.25 if two languages share both the most aggregate and same second- linguistic tree level, for instance Germanic versus Slavic languages;

all four levels of linguistic tree family, e.g. Scandinavian East (Danish, Norwegian and Swedish).. The index of linguistic proximity equal to 1 for a common language in two countries[8]. The linguistic index is based on information from Ethnologue, and is described in a greater detail in the Appendix section.

Many countries have more than one official language and among those one is the most widely used. To construct our first index of linguistic proximity we use the language most extensively used in the country. As part of the robustness analyses, we extend the set of linguistic measures to include an index that takes into account the existence of multiple official languages and we compute the index at the maximum proximity between two countries using any of those languages. The literature has shown that migrants from different linguistic backgrounds self-select to different areas within destination countries with multiple languages according to the most widely used language in each area. Chiswick and Miller (1995), one of the most prominent examples of this line of research, show how migrants to Canada self-select to the province whose language is closer to their own because that enhances their labor market returns.

In addition, we also employ a linguistic proximity measure proposed by Dyen et al. (1992), a group of linguists who built a continuous index between zero and 1000 of the distance between Indo-European languages based on the similarity of samples of words from each language. This way we are able to build a matrix that contains a continuous measure of proximity between any pair of languages from our destinations-source pairs. This should provide a better adjusted measure of proximity that the standard dummies used in most the literature. Nonetheless, the sample size in specifications containing the Dyen variable is severely reduced since only countries with Indo-European languages are included. To account for the diversity of languages in both the country of origin and destination we use a couple of indices from Desmet et al. (2011) that measure diversity: fractionalization and polarization. Desmet et al. (2011) use linguistic trees, describing the

[8] We have tried alternative indices attaching different weights to measure the relative importance of the extent of family of languages shared. The weights were based on coefficients to simple dummies for a belonging in a common linguistic family at different linguistic tree levels in migration regressions without any other controls. In particular, we set weights equal to 0.27if two languages are only related at the most aggregated linguistic tree level, e.g. Indo-European languages ; 0.45 if two languages belong to the same first and second- linguistic tree level, e.g. Germanic languages; 0.8 if two languages share the same first up to third linguistic tree level, e.g. Germanic North languages; and 0.9 if both languages share all four levels of linguistic tree family, e.g. Scandinavian East (Danish, Norwegian and Swedish), and 1 for a common language in two countries. The results from the alternative index are similar to the main results using the linguistic proximity index described in above, although the effects are even larger when using the alternative index. It is worth noting that the weights are similar to values from log transformation of the original linguistic proximity index, which is used in all our main models. We run also models with simple dummies to indicate a common linguistic family at different linguistic tree levels. The second set of results is discussed later in the paper adn result tables are available from the authors upon a request.

genealogical relationship between the entire set of 6,912 world languages, to compute measures of fractionalization and polarization at different levels of linguistic aggregation. A complete discussion about the measures can be found in their paper.

The linguistic fractionalization index computes the probability that two individuals chosen at random will belong to different linguistic groups and the index is maximized when each individual belongs to a different group. For i(j) = 1....N (j) groups of size si(j), where j = 1...J denotes the level of aggregation at which the group shares are considered[9], linguistic fractionalization is calculated as:

$$ELF(j) = 1 - \sum_{i(j)=1}^{N(j)} [si(j)]^2$$

Linguistic polarization, in contrast, is maximized when there are two groups of equal size. So if a country A consists of two linguistically different groups that are of the same size and country B has three linguistic groups of equal size, then country B is more diverse, but less polarized than A (Desmet et al. 2011).

We use the polarization measure from Desmet at al. (11 that is derived from Montalvo and Reynal-Querol (2005):

$$Pol(j) = 4 \sum_{i(j)=1}^{N(j)} [si(j)]^2 [1 \quad si(j)]$$

In addition we use three more measures from Desmet et al. (2009), GI fractionalization[10] and ER polarization indexes[11], which control for the distances between different linguistic groups in addition to their shares in the population, and PH peripheral heterogeneity index, which can be seen

[9] Even though Desmet et al. (2011) calculate these indices for 15 different levels of aggregation, in the paper we only use their measures at the 4th level of aggregation of linguistic families available in the linguistic classification of Ethnologue (e.g. German vs. English). The implied diversity of the index changes somewhat as the level of linguistic aggregation varies. Desmet et al. (2011) state in their paper that "When measured using the ELF index, the average degree of diversity rises as the level of aggregation falls, as expected. When measured using a polarization index, diversity falls at high levels of aggregation, and plateaus as aggregation falls further. (p.10)".

[10] The GI index was proposed by Greenberg (1956). It computes the population weighted total distances between all groups and can be interpreted as the expected distance between two randomly selected individuals. It is essentially a generalization of ELF, whereby distances between different groups are taken into account. Note that for this index the maximal diversity need not be attained when all groups are of the same size because it also depends on the distance between those group

[11] ER index is a special case of the family of polarization indices started by Esteban and Ray (1994) that controls for distances between linguistic groups.

as an intermediate index between fractionalization and polarization as it takes into account the distance between the center and the peripheral groups, but not between the peripheral groups themselves. Desmet et al (2009) define the distances by the number of potential linguistic branches that are shared between the languages of two groups.

Further, we include the number of indigenous languages in the country obtained from Ethnologue in order to account for the intensity of multilingualism. [12]

4.3 Econometric Approach

We first estimate the model in equation (6) by OLS starting from parsimonious to full specifications. All specifications contain a time trend variable[13] and have "robust" Hubert/White/sandwich standard errors clustered at each pair of destination and source country in order to acknowledge possible heteroscedasticity. Additionally most models contain country of destination and country of origin fixed effects. In the context of international migration research, the question of whether to account for destination- and origin-country specific effects, c_j and c_i separately or whether to include pair of countries specific effects, c_{ij} comes up regularly. Destination and origin country fixed effects might capture unobserved characteristics of immigration policy practices in each destination country, as well as climate, openness towards foreigners or culture in each country, among other things. On the other hand, pair-wise fixed effects might capture (unobserved) traditions, historical, and cultural ties between a particular pair of destination and origin countries, as well as bilateral immigration policy schemes between those countries. However, since the main focus of the paper is on the effect that linguistic and cultural proximity have on migration, and the pair-wise fixed effects would be collinear with the variables of interest, our preferred specification includes separate destination and origin country fixed effects with clustered standard errors on the level of pair of countries

Given the nonnegative nature of the data and its non-normal distribution across the sample characterized by both a relative large amount of small numbers (dispersion skewed to the left), but

[12] In separate analyses available upon request we have used another index, which limits the number of languages at the linguistic tree level 2 to those spoken by a minimum of 5% of a country's population. The measures on number of languages at different linguistic levels, spoken by different percentages of a country's population were graciously provided by Ignacio Ortuno-Ortin.

[13] In separate specifications, not presented in the paper, we used year dummies instead of a linear trend in order to control for common idiosyncratic shocks over the time period we analyze. The year dummies did not add much to the results; therefore we do not report the results in the tables but they are available from the authors upon request.

also quite a few large numbers, we also estimate the model by nonlinear least squares (NLS). Some previous studies on migration determinants have either used linear models with log-transformed variable or count models to fit the dependent variable data structure just described (e.g. Belot et al. 2008 used negative binomial; Simpson and Sparber, 2010 used Tobit and Poisson count models[14]). However the count models require the mean to be tied to the variance, which is problematic. Therefore we estimate the model using nonlinear least squares (NLS), where the level of migration flows is explained by the exponential of the linear combination of all log-transformed independent variables. This way we take into account the structure of the data, and at the same time the NLS does not impose any restrictions between the mean and the variance as some count models do. We estimate the gravity model in the following non-linear form:

$$m_{ijt} = e^{\beta_1 \ln s_{ijt-1} + \beta_2 D_{ij} + \beta_3 X_{it-1} + \beta_4 X_{jt-1} + \beta_5 L_{ij} + c_j + c_i + trend_t + \mu_{ijt}} \qquad (7)$$

In the linear and non-linear model specifications, (6) and (7) respectively, we partly control for the likely persistence of migration flows by including the lagged stock of foreigners, which in fact by construction consists of previous migration flows. In order to control fully for this persistence, and to separate pure “networks” effects from the persistence effects caused by the outcomes of previous periods, we add lagged dependent variable, which introduces an additional dynamics into the model. There is a substantial literature discussing the potential bias and inconsistency of estimators in fixed or random panel data models in a dynamic framework, as well as solutions to that, see e.g. Arellano-Bond (1991) and Arellano-Bover (1995). However, as in our model we control for fixed effects separately at the level of destinations and origins, and the dynamics are introduced on the level of country pairs, we do not run into these problems.

4. RESULTS

4.1 Linguistic proximity

[14] Simpson and Sparber (2010) discuss the “zero problem” in migration data. However, in our data only around 4,5 % of observations have a zero value, a percentage which is far from either the 95 % of zero values that Simpson and Sparber, (2010) faced or from the usual problems in the trade literature when estimating gravity models. We add a one to each observation of immigration flows and foreign population stocks prior to constructing emigration and stock rates, so that once taking logs we do not discard the “zero” observations.

Table 1, columns 1 to 5, shows pooled OLS estimates of different model specifications from parsimonious to full specification excluding unemployment rates[15]. All specifications contain a time trend variable and have "robust" Hubert/White/sandwich standard errors clustered at each pair of destination and source country.

The estimated coefficient for our variable of interest, the index of linguistic proximity, is significant and positive across all specifications. Thus, other things being equal, emigration flows between two countries are larger the closer their languages are. In column (1) the index of linguistic proximity on its own explains approximately 8.5% of the variance in emigration rates (adj. R-squared). The coefficient of 0.47 implies that moving to a destination with the same language as opposed to one with a linguistic proximity of 0.7 would be associated with an increase in emigration rates of around 20%. Unsurprisingly as additional controls are included in the model, the size of the coefficient shrinks in size. Column (2) contains other standard measures of pull and push factors from source and destination countries, such as GDP per capita, relative population, share of public expenditure in destinations to account for possible welfare magnet and distance. The coefficient of linguistic proximity decreases from 0.47 to around only 0.14, but continues to be highly significant. These additional socio-economic variables are clearly relevant in explaining the emigration flows since they account for more than 45% of the variance. In column (3) we add dummies for past colonial relationship between both countries as well as measures of distance between their capitals and an indicator of whether they share common borders. Countries are expected to be more tightly related and migration is expected to be less costly when they share a colonial past or are geographically close. Moreover, some former colonies may have adopted the language of their colonial power which we argue facilitates population movements between them. The coefficient of linguistic proximity, close to 0.1 in column (3), is only slightly affected by the inclusion of these measures. In addition to economic, colonial or geographic ties, part of the influx of new migrants into a country may be fuelled by a reduction in the moving cost to that particular destination driven by the existence of local networks and bidirectional information between both countries. Clearly, in column (4) the stock of immigrant for the same destination is positively and significantly associated with current migration flows. The explanatory power (adjusted R-squared) of the model increases from 57% to 88% when adding the lagged stock of immigrants, which indicates a strong role of network effects in driving international migration or some sort of historical path dependence in the

[15] The reason for showing the results without the unemployment variables is that the source country unemployment rates impose the largest restriction with respect to the number of missing observations. By excluding unemployment variables we have twice the number of observations as compared to the full model specification.

flows. The coefficient of the linguistic proximity drops to 0.03 when including the lagged stock of immigrants in column (4). Accounting for recent flows of immigrants to the country in the form of lagged dependent variable (lagged value of flows) in column (5) allows us to distinguish between the short-run and long-run effects. The short-run elasticity of the linguistic proximity is 0.007 and highly significant.

Besides the variables considered in our full model in column (5), there are other unobservable factors that shape international migration flows and that are characteristic of particular countries. To account for the unobserved country-specific heterogeneity, we add destination and origin country fixed-effects to the model in columns (6-8)[16]. The short-run coefficient of linguistic proximity in column (8) is 0.018, and remains highly significant at 1%, and the long-run elasticity is 0.081 Thus in the short-run the difference in emigration rates to France from either Zambia with a linguistic index of 0.1 or Sao Tome with a linguistic index of 0.7 and Benin that has French as an official language and a linguistic index of 1 (900% and 42% larger than Zambia's and Sao Tome's, respectively) will be in the order of either 16% or 0.75% (close to 1 percent), ceteris paribus.

In Table 2 we present results of our full model specification and include information on unemployment rates both at origin and destination countries. The number of observations decreases from approximately 27,000 to around 16,000 compared to models in Table 1 due to missing observations for source country unemployment rates. In addition, in Table 2 we analyze the stability of the results with respect to the choice of different econometric specifications discussed in section 4. In the first two columns we show OLS estimates. In columns 3 and 4 we present estimates of non linear least squares. Finally, we include destination and source country fixed effects to the OLS and NLS estimations, respectively in columns (6) and (7). When comparing the pooled OLS results with the NLS results and the panel models that include fixed effects for destination and source countries, the overall impression is that the sign and statistical significance of the estimated coefficients for the linguistic proximity index are quite robust across the different specifications. However, the absolute sizes of the coefficients of the linguistic proximity are generally much larger when using NLS, both with and without including destination and source country specific effects. The short-run elasticities for the index of linguistic proximity in the linear fixed-effects model in both column (8) in Table 1 and column (5) in Table 2, which include the exact same variables except for

[16] This is our preferred specification also from the statistical point of view. Besides losing our variables of interest, including instead fixed effects for each pair of countries would imply many additional parameters to be estimated and would be highly demanding for our dataset.

unemployment rates, are remarkably close, 0.018 and 0.017 respectively, despite the large reduction in the sample size.

Turning our attention to the other control variables included in the models, the coefficients of emigration rates from the previous year are always positive and highly significant indicating continuity in the direction of migration flows. The stock of immigrants from the same origin at a given destination is also positively associated with larger flows but the size of the estimated coefficient decreases substantively when the lag of the dependent variable is included. Results of the linear models with lagged dependent variable in Tables 1 to 2 indicate that a 10% increase in the stock of migrants from a certain country is associated with an increase from 1.3% to 0.8% in the emigration rate from this country, ceteris paribus.

All models contain measures of pull and push economic factors from source and destination countries, such as GDP per capita and unemployment rates as well as the share of public expenditure in destinations to account for the possibility that they act as a welfare magnet. Implied emigration rates to countries with high GDP per capita are substantial in all estimates in Tables 1 and 2, except for NLS specifications without country FE that are highly unstable and where the coefficient for GDP per capita at destination flips its sign. Once unemployment rates in countries of origin and destination are controlled for in Table 2, emigration rates are weaker from relatively richer countries of origin. Coefficients for unemployment rates in both countries are quite unstable though, in general, they point to emigration rates that are significantly higher toward countries with relatively high unemployment rates, other things being the same. This result, even if apparently surprising, may be explained by the relatively high unemployment rates experienced in many European countries during this period as compared to other areas of the OECD coupled with their comparatively large welfare states. Nonetheless country fixed-effects and time trends as well as the measure of public social expenditure should be already capturing some of those differences. The increased mobility of labor within EU countries during these last decades as barriers were dismantled may also be part of the explanation. Surprisingly, public social expenditure is inversely related to emigration rates. At any rate social expenditures would only be relevant for migrants as long as they are entitled to receive them.[17] Population ratio enters positively in pooled OLS models but it either shifts its sign in Table 1 or becomes insignificant in Table 2 when fixed effects are

[17] This is something we plan to investigate further in a separate paper.

included.[18] Distance is clearly significantly associated with weaker emigration flows in the majority of specifications. Colonial past is significantly associated with stronger emigration flows in all models (except in NLS without fixed effects), and the coefficient is predictably smaller when fixed effects are included. Emigration rates from countries with more restrictive political rights are significantly smaller in most specifications. Barriers to migration may be associated with restricted political rights in some origin countries. Civil rights do not seem to be as relevant to explain migration patterns. Only in Table 1, controlling for political rights, emigration rates seem to be larger in countries with fewer civil rights.

4.2 Robustness

As a part of the robustness analyses, we extend the set of linguistic proximity measures to include an index that takes into account the existence of multiple official languages in each country and employs the maximum proximity between two countries using any of those languages. Further, we run the regressions with the linguistic proximity index proposed by Dyen et al. (1992), a group of linguists who built a continuous measure of distance between Indo-European languages based on the similarity between samples of words from each language. Given that the Dyen index covers only Indo-European languages, our number of observations is reduced significantly from around 16,000 to only close to 9,000. Table 3 presents the results of the full model specification with country fixed effects. Columns (1) and (2) contain results of FE and NLSFE regressions with the linguistic index that takes into account multiple languages and columns (3) and (4) similar regressions using the Dyen index instead.

The coefficient of the linguistic proximity when using the multilanguage criteria is significant and positive[19] and the coefficient in column (1) is of the same size as that in column (4), Table 2, which contains the exact same specification with the basic index. Further, the Dyen index displays a significant positive coefficient in both econometric specifications in columns (3) to (4) and its magnitude is in fact much larger than that of the coefficients estimated with our basic linguistic

[18] Destination population alone, i.e. not a ratio of destination and origin population, has a large statistically significant positive coefficient in OLS regressions. However, as population does not change much over time, the coefficient becomes significantly negative in FE regressions. The choice of population measure, whether population of destination enters alone or as a ratio, does not influence other coefficients of interest. If anything the linguistic proximity index coefficient increases a bit both in OLS and NLS models with population in levels. The R-squared of the regressions is fairly similar.

[19] The coefficients for linguistic proximity are statistically insignificant in regressions without fixed effects. Tables containing results of OLS and NLS regressions without FE are available from the authors upon a request.

proximity measure shown in Tables 1 and 2. There are two potential explanations for the particular strength of this result. First, the sample is restricted to likely more homogeneous countries, since it excludes those source or destination countries with non-Indo-European languages. Second, the Dyen index allows for greater variance across country-pairs since it measures more continuously the proximity between languages than the other indicators in the paper. The magnitude of the coefficient, 0.04 in the fixed effects model, is non-negligible. For example, the difference in emigration rates to an English speaking country from Nepal (with a Dyen of 157 with respect to English) as compared to those from Zambia (with a score of 1000) should be around 21%. The difference between migrants from either Argentina (with an index of 240) or Austria (with an index of 578) with respect to someone from Zambia (with a Dyen index over 300% and 73% larger than that of Argentina and Austria) should be in order of 12.7% and 3% respectively. In separate estimates we have used the Dyen index and attached a zero value for the pairs of countries in which one language belongs to the family of Indo-European languages and the other does not[20]. The estimated coefficient on the index is, not surprisingly lower in value in full sample specifications than when the sample is restricted to indo-European countries, but it still remains significant in OLS regressions, though positive insignificant in FE and NLS specifications.

In separate models not presented here, both the estimated coefficients of the index of linguistic proximity and their significance are larger when using a sample restricted to countries with Indo-European languages than in the general sample. A possible explanation of this finding is that the linguistic index may measure better the real gap between languages within that group of countries; but the discontinuity implied by the index between Indo-European and non-Indo-European countries is excessively strict.

As another robustness analysis we run regressions with a set of dummies that indicate whether the two languages share the same linguistic family separately for each level of the linguistic tree and also a dummy that indicates when the same language is spoken in the two countries in order to depict non-linearities of the linguistic proximity index (if any). The results of FE and NLSFE regressions are presented in Table 3b, columns (1) to (5) and (6) to (10), respectively. We observe that dummies for all levels of the linguistic family tree - except for the most aggregated (Indo-European vs. Uralic) – display a significant positive coefficient, with the largest in size being the one corresponding to the fourth level of linguistic tree family e.g. Scandinavian West vs.

[20] In order to avoid the problem with zeros in logarithmic specifications, we again added 0.01 to the Dyen values.

Scandinavian East), and the second largest the one for the dummy that denotes that the same languages are spoken in the two countries[21].

Finally, one possible critique of the linguistic proximity index can be that it captures cultural proximity between countries. In order to separate the effects of language and culture, we include a couple of measures of the genetic distance between populations of both countries in our regressions. These indices are based on the work by Cavalli-Sforza, Menozzi, and Piazza (1994) and have been already been employed in other contexts to study, for example, cross-country differences in development (Spolaore and Wacziarg 2009). The first index (dominant) measures, for each pair of countries, the distance between the ethnic groups with the largest shares of population in each country. As the genetic index increases the larger are the differences between two populations. It takes a zero if the distributions of alleles in both populations are identical. The second index (weighted) takes into account within-country subpopulations that are genetically distant and calculates the distance between both countries by taking into account the difference between each pair of genetic groups and weighting them by their shares. Since both genetic distance data and data on shares of each genetic group are not available for all cases the sample is slightly smaller when this index is employed. The interpretation of this index: expected genetic distance between two randomly selected individuals, on from each country.

Results are presented in Table 3c. The first columns (1) to (4) show the coefficients for both measures of genetic distance in linear and non-linear full specifications that include fixed effects. The coefficient for the weighted index is negative indicating weaker migration flows when the genetic distance is larger, though the coefficient is only significant in the NLSFE model. For the dominant genetic distance the coefficient is also negative and significant in the non-linear model in column (4). Columns (5) through (8) in Table 3c show that our linguistic proximity index is robust to the inclusion of both measures of genetic distance. This suggests that language on its own affects migration costs beyond any ease derived from moving to a destination where people may look or be culturally more similar to the migrant. Coefficients for the linguistic proximity variable are remarkably similar to those in Table 2.

4.3 The Role of Widely Spoken Languages

[21] We have also run models that include dummies for all levels of linguistic tree families in one regression. However due to high multicollinearity among those dummies, coefficients were not surprisingly insignificant on their own in most models. Results are available upon a request.

Our linguistic proximity index does not take into account completely the importance of the use of some widely spoken Indo-European languages (particularly English) in the media (TV, music) internet, business or everyday life and the high frequency of English as a choice of second language in schools. Therefore as a further step in our analysis we divide our data by non-English and English speaking destinations in order to examine the role of English as a widely spoken language. If there is some “proficiency” advantage from knowing English as a second language, we expect that the linguistic proximity between native languages should matter more in the sample of non-English speaking destinations than for the rest. Results in Table 4 confirm this hypothesis. The linguistic proximity is a strong predictor of emigration rates in the sample of non-English speaking destinations. However, among English-speaking destinations the linguistic proximity is insignificant. This gives support to the hypothesis that people tend to migrate to destinations with a widely spoken language no matter how far linguistically their mother languages are from the language of the country of destination. There may be two different forces driving this result. First, even if they do not regularly speak it at home, many migrants would have previous knowledge of a widely spoken language taught at schools and used in the internet and movies, particularly English (see special Eurobarometer study on languages by European Commission, and Pytlikova 2006). Second, foreign language proficiency is an important part of human capital in the labor market of source countries, see e.g. European Commission (2002) on language proficiency as an essential skill for finding a job in home countries. Thus learning/practicing/improving the skills of “widely spoken” language in the “native” countries serve as a pull factor especially for temporary migrants who may take this skill back home.

In additional models available upon request we have also included measures of the number of computers per capita in the country to calculate the access to information about countries, or to infer exposure to English or other languages though internet and media use. All results remain unchanged.

4.4 Linguistic Diversity and Polarization

Table 5 includes a set of measures of the linguistic fractionalization and polarization of sending and receiving countries as defined in section 4. Each one of the boxes corresponds to a different model that, in addition to the two coefficients presented in the table, also includes covariates for linguistic proximity, network, economic conditions, distance and a time trend. Each model is first estimated by OLS and then with the nonlinear specification. None of the models includes fixed effects because the diversity and polarization indices are constant for each country. The first row includes two measures of diversity of languages both at origin and at destination. The ELF measure presented in the first two columns measures the fractionalization of languages at the level 4 of the linguistic tree and is obtained from Desmet (2011). The following two columns use indices measuring polarization at the 4th level in the linguistic tree of Ethnologue. Results from both fractionalization and polarization indices are fairly similar. Coefficients for the diversity of languages at destination are negative and highly significant in all specifications. Ceteris paribus, the higher the linguistic diversity at destination, the smaller the migration flows. The mechanism behind this finding is subject of speculation but may be related to fear from migrants that adaptation will be costly when not only one but more languages need to be learnt. On the other hand one could have expected that places with a tradition of linguistic diversity could be welcoming to people with a different linguistic background. Conversely, the flows of migrants from countries with high linguistic diversity are larger than others (only in the OLS estimates). Again, the explanations for the finding stretch from either diversity bolstering internal conflict and acting as a push factor for migrants to alternatively diversity viewed as an asset that facilitates language acquisition at destination and lowers migration costs.

The second row in Table 5 includes regressions with diversity indices, both at destination and at origin, which take into account the linguistic distance between each pair of languages. The fractionalization is represented by the GI index from Desmet (2009), which takes into account the actual distance of languages and not only the particular linguistic family to which they belong as the ELF indices do. The polarization is now measured by ER index (of the family of polarization measures started by Esteban and Ray), which takes into account not only the different number of languages and their share of speakers but also the linguistic distance between each pair of languages. Interestingly, once we control for linguistic distances the coefficients to fractionalization and polarization differ. In particular, the coefficients to the ER polarization index become larger in absolute terms, while coefficients to the GI fractionalization index become smaller.

... this supports are hypothesis that people do not want to invest into two very different languages..A more deeply polarized linguistic environment at destination seems to deter migration flows, other things being the same. Conversely, more polarized societies seem to significantly push larger number of people in the search of a new life elsewhere though the coefficients are somewhat unstable across estimation methods. We also run regressions with PH peripheral diversity index studied by Desmet et al. (2005), which also account for distances but not among all linguistic groups as in the previous indexes, but between the center and the peripheral groups. Not surprisingly the coefficients to the PH index lie somewhat between the coefficients of GI fractionalization and ER polarization.

Finally an additional variable measuring the linguistic richness of both country of origin and destination are presented in the third row of Table 6: The total number of indigenous languages at the linguistic three level 2 spoken by at least 5 % of the population at the country of destination are consistently negatively associated with the dimension of flows. However, results for the source country are inconclusive –coefficients shift once again from significantly positive in OLS to negative in NL models.

5. CONCLUSIONS and FURTHER STEPS

Fluency in the language of the destination country plays an important role in the transfer of human capital of migrants to a foreign country and generally it reduces migration costs and increases the rate of success of immigrant at the destination country’s labor market. Previous research has already shown that sharing a language is associated with larger population movements across countries. In this paper we use data on immigration flows and stocks of foreigners in 27 OECD destination countries from 130 source countries for the years 1985–2006 to study the role of language in shaping international migration in more detail. In addition to standard covariates from gravity models (e.g. income per capita, unemployment, distance, colonial past, welfare expenditures), we include a set of indices of language proximity to study their association with the observed flows.

We find that emigration rates are higher among countries whose languages are more similar. The result holds both for the analysis of the proximity between the most used language in each country as well as to the minimum distance between any of the official languages in both countries. Among countries with Indo-European languages this result is highly robust to the use of an alternative continuous distance measure developed by a group of linguists. Further, our linguistic proximity index is robust to the inclusion of genetic distance, which suggests language itself affects migration costs beyond any ease derived from moving to a destination where people may look or be culturally more similar to the migrant. When splitting the sample between English and non-English speaking destinations, linguistic proximity matters significantly for the latter group. A likely higher English proficiency of the average migrant may diminish the relevance of the linguistic proximity to English speaking destinations. Finally, destinations that are more diverse and polarized linguistically attract fewer migrants; whereas more linguistic polarization at origin seems to act as a push factor.

Of course the current paper has some limitations. First, the linguistic proximity between countries is constructed here as symmetric. It is possible that even if two countries belong to the same branch in the linguistic tree, the easiness of learning each one of the languages by an individual of the other country may not be exactly the same (e.g. the complexity of grammar and phonemes varies within similarly homogenous groups such as romance languages). Second, our indices are unable to capture the familiarity of migrants to other widely spoken languages (than English) that may have been learn in school or though media use. The extent of dubbing varies across the world and obtaining a good measure of the exposure of residents in each country to original movies or TV shows would prove very interesting. Third, positive selection of migrants that may imply over the average knowledge of second languages could also matter. Fourth, diversity at origin may be confounding of violence or conflict. In a separate paper we will explore these mechanisms.

Despite some of these potential shortcomings in interpretation and data availability, this is, to our knowledge, the first paper that approaches the relationship between migration rates and language using new linguistic proximity measure and using information on migration for such a large set of origin and destination countries that spans for three decades. Further, it uses multiple measures of language for any pair of destination-origin countries.

<u>References:</u>

Arellano, M. and S. Bond (1991): “Some Tests Of Specifications for Panel Data: Monte Carlo Evidence And An Application To Employment Equations” *Review of Economic Studies*, Vol.58, No.2, pp. 279 – 299.

Arellano, M and O. Bover (1995): “Another Look at the Instrumental Variable Estimation of Error Components Models” *Journal of Econometrics,* Vol. 68, pp. 29-51.

Adsera, A. and B. R. Chiswick, (2007). Are There Gender and Country of Origin Differences in Immigrant Labor Market Outcomes across European Destinations? *Journal of Population Economics*, Vol. 20 (3), 495-526.

Bauer, T. K., I. N. Gang and G. Epstein (2007), The Influence of Stocks and Flows on Migrants' Location Choices. *Research in Labor Economics,* Vol. (26), 199-229.

Bauer, Thomas, and Gil Epstein, and Ira Gang, (2005) "Enclaves, Language, and the Location Choice of Migrants," *Journal of Population Economics*, Vol. 18(4), pages 649-662,

Beine, M., Docquier, Frederic and C. Ozden (2011): “Diasporas”, forthcoming in *Journal of Development Economics* 95, 30-41.

Belot, M. and S. Ederveen (2010): “Cultural and Institutional Barriers in Migration between OECD Countries”, forthcoming in *Journal of Population Economics.*

Belot, M. and T. Hatton (2011) Skill Selection and Immigration in OECD Countries, *Scandinavian Journal of Economics*, forthcoming

Bleakley, H. and A. Chin (2004): “Language Skills and Earnings: Evidence from Childhood Immigrants”, *Review of Economics and Statistics* 84 (2), 481-496.

Bleakley, H. and A. Chin (2010): “Age at Arrival, English Proficiency, and Social Assimilation among US Immigrants”, *American Economic Journal: Applied Economics 2(1), 165-192.*

Borjas, G. J. (1999). *Heaven’s Door: Immigration Policy and the American Economy*. Princeton, NJ: Princeton University Press.

Borjas, G. J. (2003). "The Labor Demand Curve Is Downward Sloping: Reexamining The Impact of Immigration on The Labor Market," *The Quarterly Journal of Economics*, MIT Press, vol. 118(4), pages 1335-1374.

Boyd, M. (2010) “Language at Work: The Impact of Linguistic Enclaves on Immigrant Economic Integration”, paper presented at PAA meeting, Dallas.

Card, David (2005) “Is the New Immigration Really so Bad?” *Economic Journal,* 115, pp. 300-323.

Cavalli-Sforza, Luigi L., Paolo Menozzi, and Alberto Piazza (1994) *The History and Geography of Human Genes* (Princeton, NJ: Princeton University Press).

Chiswick, B. R., (1991): “Speaking, Reading and Earnings among Low-Skilled Immigrants”, *Journal of Labor Economics*, Vol. 9 (2), 149-170.

Chiswick, B. (1996): “The Economic Consequences of Immigration: Application to the United States and Japan” in Weiner M. and T. Hanami, eds: Temporary Workers and Future Citizens? Japanese and U.S. Migration Policies. New York: New York University Press, pp. 177 - 208.

Chiswick, B. R., and T. J. Hatton (2002). ‘International Migration and the Integration of Labor Markets’, in M. Bordo, A. M. Taylor, and J. G. Williamson (eds), *Globalization in Historical Perspective*. Chicago: University of Chicago Press.

Chiquiar, Daniel, and Gordon Hanson (2005): “International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the U.S.,” *Journal of Political Economy* 113, pp. 239–281.

Chiswick, B.R. and P.W. Miller (1995), “The Endogeneity between Language and Earnings,” *Journal of Labor Economics*, 13 (2), pp. 246-288.

Chiswick, B.R. and P.W. Miller (2002), Immigrant Earnings: Language Skills, Linguistic Concentrations and the Business Cycle” , *Journal of Population Economics*, 15(1) January 2002, pp. 31-57.

Chiswick, B.R. and P.W. Miller (2007), Computer Usage, Destination Language Proficiency and the Earnings of Natives and Immigrants,” *Review of the Economics of the Household*, 5 (2), June 2007, pp. 129-157.

Chiswick, B.R., and P.W. Miller (2010), “Occupational Language Requirements and the Value of English in the US Labor Market.” *Journal of Population Economics*, 23(1): 353–372.

Clark, X., T.J Hatton and J. G Williamson, (2007): "Explaining U.S. Immigration, 1971-1998," *The Review of Economics and Statistics*, MIT Press, vol. 89(2), pages 359-373

D’Amuri F., Ottaviano I.P. Gianmarco and G. Peri (2010): “The Labor Market Impact of Immigration in Western Germany in the 1990s” *European Economic Review*, **Vol. 54 (4),550-570**

Desmet, K., I. Ortuño-Ortín and R.Wacziarg (2011), The Political Economy of Ethnolinguistic Cleavages, *Journal of Development Economics*

Desmet, K., I. Ortuño-Ortín and S. Weber (2009), "Linguistic Diversity and Redistribution", *Journal of the European Economic Association*, vol. 7, no. 6, December.

Duclos, J-Y, J.M. Esteban, and D. Ray (2004), “Polarization: Concepts, Measurement, Estimation”, *Econometrica* 72, 1737-1772.

Dustmann, Christian. (1994). “Speaking Fluency, Writing Fluency and Earnings of Migrants.” *Journal of Population Economics* 7, pp. 133–56.

Dustmann, Ch. and A. van Soest (2001): “Language Fluency and Earnings: Estimation with Misclassified Language Indicators.” *The Review of Economics and Statistics*, Vol. 83, No. 4, pp. 663-674.

Dustmann, Ch. and A. van Soest (2002): “Language and the Earnings of Immigrants.” *Industrial and Labor Relations Review”* 55 (3), pp.473–492.

Dustmann, C. and F. Fabbri, (2003): “Language Proficiency and Labour Market Performance of Immigrants in the UK”, *Economic Journal*, Vol. 113, 695-717.

Dyen I., Kruskal J.B. and P. Black (1992): “An Indo-European classification: A lexicostatistical experiment”. Transactions of the American Philosophical Society 82/5. Philadelphia.

Esteban, J. M., and D. Ray (1994), "On the Measurement of Polarization, *Econometrica*, vol. 62, no. 4, pp. 819-851.

Esteban, J.M., and D. Ray, (2006): “Polarization, Fractionalization and Conflict,” mimeo.

Esteban, J.M., and D. Ray, (2010): “Linking Conflict to Inequality and Polarization”, *American Economic Review.*

Ethnologue: Languages of the World, 14th edition. http://www.ethnologue.com/web.asp

European Commission (2002, 2003): Candidate Countries Eurobarometer (CCB) http://europa.eu.int/comm/public_opinion/cceb_en.htm and Special Eurobarometer on “Europeans and their languages”: http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_sum_en.pdf

Falck, Oliver & Heblich, Stephan & Lameli, Alfred & Suedekum, Jens, (2010). "Dialects, Cultural Identity, and Economic Exchange," IZA Discussion Papers 4743, Institute for the Study of Labor (IZA).

Fearon, James D. (2003): “Ethnic and Cultural Diversity by Country, “*Journal of Economic Growth* 8, 195-222.

Fearon, James D., and David D. Laitin,(2003) “Ethnicity, Insurgency, and Civil War,” *American Political Science Review* 97, 75–90.

Fertig, M., Schmidt, C.M., (2000). “Aggregate-level migration studies as a tool for forecasting future migration streams”, IZA Discussion Paper 183, IZA, Bonn.

Fidrmuc, Jan & Jarko Fidrmuc, (2009). "Foreign Languages and Trade", CEDI Discussion Paper Series 09-03, Centre for Economic Development and Institutions (CEDI), Brunel University

Filer, R. (1992): “The Impact of Immigrant Arrivals on Migratory Patterns of Native Workers” in Borjas G. and R. Freeman, eds.: Immigration and the Workforce: Economic Consequences for the United States and Source Areas. Chicago University Press, Chicago, pp. 93-134.

Friedberg, Rachel M & Hunt, Jennifer, (1995). "The Impact of Immigrants on Host Country Wages, Employment and Growth," *Journal of Economic Perspectives*, vol. 9(2), pp. 23-44.

Greenberg, J. H. (1956). “The measurement of linguistic diversity.” *Language* 32, pp. 109-115.

Guiso, Luigi, Paolo Sapienza and Luigi Zingales. (2009) “Cultural Biases in Economic Exchange?.” *Quarterly Journal of Economics* 124, 3.

Hanson Gordon (2009): “The Economic Consequences of International Migration,” *Annual Review of Economics*, 1: 179-208.

Hatton, T.J and J.G. Williamson (2005): “What Fundamentals Drive World Migration?” in G. Borjas an J. Crisp (eds), *Poverty, International Migration and Asylum*, Palgrave-Macmillan.

Harris, J.R., Todaro, M.P., (1970). “Migration, unemployment and development: A two-sector analysis”, *American Economic Review* Vol. 60 (5), pp. 126–142.

Heitmueller, A. (2006): “Coordination Failures In Network Migration,” Manchester School, University of Manchester, vol. 74(6), pages 701-710

Hunt, Jennifer, 1995. "The Effect of Unemployment Compensation on Unemployment Duration in Germany," *Journal of Labor Economics*, vol. 13(1), pages 88-120, January.

Karemera, D, Iwuagwu Oguledo, V., and Davis, B. (2000) A Gravity Model Analysis of International Migration to North America, *Applied Economics*, 32, 1745-1755.

Kossoudji, S.A. (1988): “The Impact of English Language Ability on the Labor Market Opportunities of Asian and Hispanic Immigrant Men”. *Journal of Labor Economics*. Vol. 6(3), pp. 205-228.

Kovacs, A. M. & Mehler, J. (2009): “Flexible Learning of Multiple Speech Structures in Bilingual Infants.” *Science* Vol. 325, pp. 611-612.

Massey, D., J. Arango, G. Hugo, A. Kouaci, A. Pellegrino, and E. Taylor (1993):, “Theories of International Migration: *A Review and Appraisal’ Population and Development Review*, vol. 19, no. 3, pp. 431–466.

Mayda A.M. (2010): “International migration: A panel data analysis of the determinants of bilateral flows”, *Journal of Population Economics* Vol. 23(4), pp 1249-1274.

McKenzie, D. and H. Rapoport (2010): “Self-Selection Patterns In Mexico-U.S. Migration: The Role of Migration Networks”, *The Review of Economics and Statistics,* Vol 92(4), pp. 811-821.

Montalvo, J. G. and M. Reynal-Querol (2005), "Ethnic Polarization, Potential Conflict and Civil War", *American Economic Review,* vol. 95, no. 3, June, pp. 796-816.

Munshi, K., (2003) “Networks in the modern economy: Mexican migrants in the US labor market”. *The Quarterly Journal of Economics* 118 (2), 549–599.

Ortega, Francesc & Peri, Giovanni, (2009). "The Causes and Effects of International Migrations: Evidence from OECD Countries 1980-2005," NBER Working Papers 14833, National Bureau of Economic Research.

Ottaviano Gianmarco, and Giovanni Peri, (2011), “Rethinking the Effects of Immigration on Wages,” *Journal of the European Economic Association, forthcoming,*

Pedersen, P., M. Pytlikova and N. Smith (2008), Selection and Network Effects – Migration Flows into OECD Countries, 1990-2000, *European Economic Review,* Elsevier, vol. 52(7), pages 1160-1186.

Peri Giovanni, (2010) “The Effect of Immigration on Productivity: Evidence from U.S. States,” the *Review of Economics and Statistics, Forthcoming*.

Pytlikova, M (2006),”Where did Central and Eastern European Emigrants Go and Why?” unpublished manuscript.

Toomet, Ott. 2011. "Learn English, Not the Local Language! Ethnic Russians in the Baltic States." *American Economic Review*, 101(3): 526–31.

Simpson and Sparber (2010) “The Short- and Long-Run Determinants of Unskilled Immigration into U.S. States”, Colgate University Working Paper 2010-06.

Spolaore, E. and R. Wacziarg, (2009), “The Diffusion of Development” *Quarterly Journal of Economics*, 124 (2), pp. 469-530.

Sjastaad L., (1962) The Costs and Returns of Human Migration, *Journal of Political Economy* (70), 80-93

Vogler, M. and R. Rotte (2000): "The effects of development on migration: Theoretical issues and new empirical evidence," Journal of Population Economics, Springer, vol. 13(3), pages 485-508.

Zavodny, M. (1997): “Welfare and the Locational Choices of New Immigrants” *Economic Review – Federal Reserve Bank of Dallas;* Second Quarter 1997, pp. 2-10.

I. The role of linguistic distance:

Table 1: OLS and FE (destinations and origins) Estimation of migration flows from 130 countries of origin (i) to 27 OECD destination countries (j), 1985-2006.

	OLS	OLS	OLS	OLS	OLS	FE	FE	FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Log Emigration Rate	Log Emigration Rate	Log Emigration Rate	Log Emigration Rate	Log Emigration Rate	Log Emigration Rate	Log Emigration Rate	Log Emigration Rate
Log Linguistic Proximity	0.467***	0.137***	0.097***	0.031***	0.007**	0.246***	0.032	0.018***
	(0.029)	(0.024)	(0.023)	(0.012)	(0.003)	(0.036)	(0.019)	(0.006)
Log Emigration Rate_t-1	-	-	-	-	0.848***	-	-	0.778***
	-	-	-	-	(0.007)	-	-	(0.009)
Log Stock of Migrants_t-1	-	-	-	0.719***	0.096***	-	0.695***	0.131***
	-	-	-	(0.009)	(0.006)	-	(0.011)	(0.007)
Log Destination	-	2.089***	2.130***	0.501***	0.130***	1.819***	3.242***	1.196***
GDPperCapPPPj_t-1	-	(0.101)	(0.097)	(0.059)	(0.015)	(0.207)	(0.194)	(0.088)
Log Origin	-	0.480***	0.535***	-0.003	0.009**	-0.547***	-0.265***	0.009
GDPperCapPPPi_t-1	-	(0.028)	(0.037)	(0.020)	(0.004)	(0.084)	(0.086)	(0.037)
Log Destination	-	-0.672***	-0.651***	-0.162**	-0.070***	0.001	0.409***	-0.039
Public Social Expenditure_t-1	-	(0.139)	(0.134)	(0.074)	(0.014)	(0.114)	(0.113)	(0.051)
Log Population Ratio_t-1	-	0.629***	0.629***	0.130***	0.025***	1.323***	0.490***	-0.110*
	-	(0.016)	(0.016)	(0.010)	(0.002)	(0.164)	(0.146)	(0.058)
Log Distance in km	-	-0.461***	-0.331***	-0.236***	-0.053***	-0.919***	-0.380***	-0.108***
	-	(0.035)	(0.037)	(0.021)	(0.005)	(0.062)	(0.035)	(0.010)
Neighbouring Dummy	-	-	1.346***	0.017	-0.001	0.363**	-0.121*	-0.024
	-	-	(0.156)	(0.086)	(0.017)	(0.145)	(0.073)	(0.019)
Historical Past Dummy	-	-	2.034***	0.186	0.112***	1.866***	0.297***	0.099***
	-	-	(0.176)	(0.129)	(0.030)	(0.206)	(0.111)	(0.031)
Log Origin Freedom	-	-	-0.139**	0.003	-0.018*	0.054*	0.030	0.006
Political Rightsi_t-1	-	-	(0.068)	(0.040)	(0.010)	(0.032)	(0.028)	(0.013)
Log Origin Freedom	-	-	0.287***	-0.001	0.033***	-0.151***	-0.025	0.001
Civil Rightsi_t-1	-	-	(0.081)	(0.049)	(0.012)	(0.036)	(0.031)	(0.016)
Destination & Origin FE	NO	NO	NO	NO	NO	YES	YES	YES
Time Trend	0.020***	-0.027***	-0.027***	-0.005**	-0.001	0.011**	-0.058***	-0.024***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.001)	(0.005)	(0.005)	(0.002)
Constant	-44.277***	24.805***	22.564***	3.271	0.055	-41.354***	79.837***	35.927***
	(5.329)	(5.163)	(5.103)	(3.719)	(1.055)	(8.971)	(8.561)	(3.654)
Observations	*45,950*	*39,737*	*39,313*	*26,822*	*25,651*	*39,313*	*26,822*	*25,651*
Adjusted R-squared	0.085	0.540	0.573	0.877	0.960	0.778	0.912	0.962

Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Political rights and civil liberties are measured from highest (value 1) to lowest (value 7).

Table 2: OLS, NLS and FE (destinations and origins), adding unemployment rates. Estimation of migration flows from 130 countries of origin (i) to 27 OECD destination countries (j), 1985-2006.

	OLS	OLS	FE	FE	NLS	NLS	NLS FE	NLS FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Log Emigration Rate	Log Emigration Rate	Log Emigration Rate	Log Emigration Rate	Emigration Rate	Emigration Rate	Emigration Rate	Emigration Rate
Log Linguistic Proximity	0.025*	0.006*	0.034	0.017***	0.157***	0.0385**	0.0880	0.0826*
	(0.015)	(0.003)	(0.021)	(0.006)	(0.0580)	(0.0157)	(0.0830)	(0.0442)
Log Emigration Rate_t-1	-	0.864***	-	0.794***	-	0.835***	-	0.742***
	-	(0.009)	-	(0.012)	-	(0.0304)	-	(0.0522)
Log Stock of Migrants_t-1	0.733***	0.086***	0.705***	0.128***	0.701***	0.0674***	0.669***	0.0128
	(0.012)	(0.007)	(0.013)	(0.009)	(0.0495)	(0.0224)	(0.0986)	(0.0694)
Log Destination GDPperCapPPPj_t-1	0.279***	0.135***	2.972***	1.215***	-1.382***	-0.325**	0.242	0.809**
	(0.084)	(0.019)	(0.260)	(0.122)	(0.389)	(0.153)	(1.750)	(0.342)
Log Origin GDPperCapPPPi_t-1	-0.095***	-0.020***	-0.180	-0.055	-0.585***	-0.263***	-1.346**	-0.753***
	(0.029)	(0.007)	(0.110)	(0.040)	(0.143)	(0.0851)	(0.673)	(0.202)
Log Destination Public Social Expenditure_t-1	0.055	-0.078***	0.180	-0.112*	0.422	-0.0901	-2.427**	-0.936***
	(0.098)	(0.018)	(0.133)	(0.060)	(0.464)	(0.108)	(1.006)	(0.329)
Log Destination UnemplRate_t-1	-0.107***	0.044***	-0.071**	0.051***	0.0197	0.0163	0.00623	0.115
	(0.038)	(0.010)	(0.036)	(0.014)	(0.155)	(0.0747)	(0.228)	(0.0888)
Log Origin UnemplRate_t-1	0.029	0.004	0.088***	0.032**	0.0387	-0.0767*	-0.0380	-0.152*
	(0.028)	(0.006)	(0.030)	(0.014)	(0.123)	(0.0403)	(0.208)	(0.0782)
Log Population Ratio_t-1	0.116***	0.016***	0.255	-0.114	0.0682	0.0251	1.227	0.461
	(0.012)	(0.003)	(0.213)	(0.084)	(0.0456)	(0.0169)	(1.539)	(0.562)
Log Distance in km	-0.191***	-0.044***	-0.372***	-0.087***	0.0206	-0.0382	-0.323***	-0.0937*
	(0.023)	(0.005)	(0.038)	(0.010)	(0.113)	(0.0251)	(0.105)	(0.0494)
Neighbouring Dummy	0.067	0.003	-0.162**	-0.028	0.325	0.00248	-0.196	0.0242
	(0.088)	(0.016)	(0.075)	(0.018)	(0.204)	(0.0620)	(0.189)	(0.0904)
Historical Past Dummy	0.219	0.101***	0.263**	0.075**	-0.907**	0.0720	0.529**	0.312*
	(0.147)	(0.032)	(0.127)	(0.033)	(0.405)	(0.0717)	(0.254)	(0.173)
Log Origin Freedom PoliticalRi_t-1	-0.025	-0.027**	0.053	0.001	-0.105	-0.260***	0.000291	-0.380***
	(0.047)	(0.011)	(0.039)	(0.018)	(0.145)	(0.0948)	(0.172)	(0.122)
Log Origin Freedom CivilRi_t-1	-0.086	0.006	-0.060*	-0.018	-0.189	-0.0950	-0.0880	0.0407
	(0.053)	(0.013)	(0.035)	(0.019)	(0.171)	(0.0813)	(0.109)	(0.0989)
Destination and Origin FE	NO	NO	YES	YES	NO	NO	YES	YES
Time Trend	0.011***	0.003***	-0.049***	-0.021***	0.0080***	0.0032***	0.00381	-0.000800
	(0.003)	(0.001)	(0.006)	(0.003)	(0.00231)	(0.00111)	(0.0112)	(0.00264)
Constant	-26.473***	-6.475***	66.969***	30.846***	-	-	-	-
	(4.822)	(1.334)	(10.235)	(4.461)	-	-		-
Observations	*16,814*	*16,221*	*16,814*	*16,221*	*16814*	*16221*	*16814*	*16221*
Adjusted R-squared	0.872	0.963	0.913	0.965	0.679	0.908	0.826	0.919

Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table 3a: Robustness checks – using Dyen index and linguistic distance controlling for all possible language as alternative measures for linguistic proximity. Controlling for genetic distance in order to separate cultural and linguistic proximity. Estimation of migration flows from 130 countries of origin (i) to 27 OECD destination countries (j), 1985-2006. FE and NLSFE.

	FE	NLSFE	FE	NLSFE	FE	NLSFE
VARIABLES	(1) Log Emigration Rate	(2) Emigration Rate	(3) Log Emigration Rate	(4) Emigration Rate	(5) Log Emigration Rate	(6 Emigration Rate
Log Linguistic Proximity_All	0.017**	0.122**	-	-	-	-
	(0.007)	(0.0533)	-	-	-	-
Log Dyen	-	-	0.041***	0.174*	-	-
	-	-	(0.012)	(0.0952)	-	-
Log Linguistic Proximity	-	-	-	-	0.022***	0.0874*
	-	-	-	-	(0.007)	(0.0480)
Genetic Distance Dominant					0.000**	0.000091
					(0.000)	(0.00015)
Constant	30.617***	-	23.818***	-	30.761***	-
	(4.474)	-	(5.480)	-	(4.466)	-
Observations	*16,221*	*16221*	*8,900*	*8900*	*16221*	*16221*
Adjusted R-squared	0.965	0.918	0.969	0.916	0.965	0.919

Notes: Controls included: lagged dependent variable, networks, economic variables, distance variables, time trend and destination and origin country fixed effects. Robust standard errors clustered on country pairs level, *** p<0.01, ** p<0.05, * p<0.1

Table 3b: Robustness checks – using linguistic family dummies as alternative measures for linguistic proximity. Estimation of migration flows from 130 countries of origin (i) to 27 OECD destination countries (j), 1985-2006. FE and NLSFE.

	FE	FE	FE	FE	FE	NLSFE	NLSFE	NLSFE	NLSFE	NLSFE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
same1	-0.000	-	-	-	-	0.346** *	-	-	-	-
	(0.026)	-	-	-	-	(0.108)	-	-	-	-
same2	-	0.038***	-	-	-	-	0.137	-	-	-
	-	(0.014)	-	-	-	-	(0.0886)	-	-	-
same3	-	-	0.050***	-	-	-	-	0.179**	-	-
	-	-	(0.014)	-	-	-	-	(0.0848)	-	-
same4	-	-	-	0.069***	-	-	-	-	0.0604	-
	-	-	-	(0.018)	-	-	-	-	(0.0736)	-
SameLanguage	-	-	-	-	0.056**	-	-	-	-	0.0874
	-	-	-	-	(0.028)	-	-	-	-	(0.130)
Constant	31.188***	30.796***	31.017***	30.653***	31.000***	-	-	-	-	-
	(4.465)	(4.477)	(4.471)	(4.472)	(4.470)	-	-	-	-	-
Observations	*16221*	*16221*	*16221*	*16221*	*16221*	16221	16221	16221	16221	16221
Adjusted R-squared	0.965	0.965	0.965	0.965	0.965	0.919	0.919	0.919	0.919	0.919

Notes: Controls included: lagged dependent variable, networks, economic variables, distance variables, time trend and destination and origin country fixed effects. Robust standard errors clustered on country pairs level, *** p<0.01, ** p<0.05, * p<0.1

Table 4: The role of English as widely spoken language – division by non-English and English speaking destinations.

	Non-English speaking destinations		English speaking destinations	
	(1)	(2)	(3)	(4)
VARIABLES	Log Emigration Rate	Emigration Rate	Log Emigration Rate	Emigration Rate
	FE	NLSFE	FE	NLSFE
Log Linguistic Proximity	0.031***	0.150***	0.018	-54.54*
	(0.008)	(0.0535)	(0.202)	(32.00)
Constant	26.731***	-	49.554***	-
	(4.787)	-	(15.939)	-
Observations	*13770*	*13770*	*2451*	*2451*
Adjusted R-squared	0.962	0.922	0.932	0.925

Notes: Controls included: lagged dependent variable, networks, economic variables, distance variables, time trend and destination and origin country fixed effects. Robust standard errors clustered on country pairs level, *** p<0.01, ** p<0.05, * p<0.1

Table 5: Linguistic Diversity in destinations and origins, Estimation of migration flows from 130 countries of origin(i) to 27 OECD destination countries(j), 1985-2006

	(1)	(2)	(3)	(4)
Linguistic diversity	OLS Log Emigration Rate	NLS Emigration Rate	OLS Log Emigration Rate	NLS Emigration Rate
Measured by: In:	LogELF – a diversity index without distances		LogPOL- a polarization index without distances	
Destination	-0.022*** (0.005)	-0.137*** (0.0449)	-0.021*** (0.005)	-0.140***(0.0465)
Origin	0.014*** (0.003)	-0.0136 (0.0128)	0.014***(0.004)	-0.0155 (0.0125)
Observations	*16221*	*16221*	*16221*	*16221*
Adj. R2	0.963	*0.910*	0.963	0.910
	LogGI - a diversity index with distances		LogER - a polarization index with distances	
Destination	-0.014** (0.006)	-0.116** (0.0526)	-0.026*** (0.008)	-0.165** (0.0724)
Origin	0.009** (0.004)	-0.0129 (0.0129)	0.021*** (0.006)	-0.00922 (0.0164)
Observations	*14815*	*14815*	*14815*	*14815*
Adj. R2	*0.964*	*0.909*	0.964	0.909
Linguistic diversity in:	LogPH– peripheral diversity index		LogNoLangT2P5j - number of languages	
Destination	-0.016***(0.006)	-0.123** (0.0579)	-0.086***(0.013)	-0.371***(0.133)

Origin	0.014*** (0.005)	-0.00864 (0.0131)	0.028***(0.010)	-0.0765*(0.0445)
Observations	*14815*	*14815*	*16221*	*16221*
Adj. R2	0.964	0.909	*0.963*	*0.910*

The table shows results of full model specification, i.e. we control for: lagged dependent variable, networks, linguistic proximity, network, economic, distance variables and time trend. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Appendix section

Table A: Descriptive statistics

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        year |     77814      1995.5     6.34433       1985       2006
      flowij |     47438    1445.654    8706.034          0     946167
     stockij |     39940    26097.27    198549.2          0   1.15e+07
       pop_j |     77814    3.30e+07    5.45e+07     241000   2.98e+08
       pop_i |     73602    4.46e+07    1.39e+08     103852   1.31e+09
-------------+--------------------------------------------------------
   gdpPPP05_j |     76504    25989.79    9018.388   7567.728   70762.47
   gdpPPP05_i |     67122    9896.994    10947.27    244.326   70762.47
       psepj |     58817    21.10004    4.788428         11       36.2
      unpl_j |     71395    7.661596    4.149071       1.48      23.88
      unpl_i |     37665    8.366122    5.046069         .3      31.22
-------------+--------------------------------------------------------
     freepri |     72522    3.690246    2.240432          1          7
     freecri |     72521    3.788117    1.943387          1          7
      distij |     76604    6438.097    4366.771       60.2      19900
   neighbour |     77814    .0359061    .1860573          0          1
      colony |     77814    .0245971    .1548948          0          1
-------------+--------------------------------------------------------
      index2 |     77814    .127594     .2237484          0        1
      elf_1i |     76032    .1434539    .1697655          0      .6466
      elf_4i |     76032    .2842711    .2329973          0       .857
      pol_1i |     76032    .2648453    .3028694          0      .9976
      pol_4i |     76032     .413382    .2898045          0      .9911
-------------+--------------------------------------------------------
      elf_1j |     77814       .0557      .06229          0      .2545
      elf_4j |     77814    .1870296    .1606463      .0109      .5783
      pol_1j |     77814    .1085148    .1196935          0      .4736
      pol_4j |     77814    .3280444    .2635794      .0218       .923
          HI |     77814     .091719    .1204963          0          1
-------------+--------------------------------------------------------
     HIflows |     77814    .1044508    .1301712          0          1
```

I. List of destination countries,

Australia, Austria, Belgium, Canada, Czech Republ, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Luxembourg, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Repub, Spain, Sweden, Switzerland, United Kingd, United States

II. List of countries of origin:

Afghanistan, Albania, Algeria, Angola, Argentina, Australia, Austria, Azerbaijan, Bangldesh, Belarus, Belgium, Benin, Bolivia, Bosnia Hercegovina, Brazil, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Cape Verde, Chad, Chile, China, Taiwan, Colombia, Côte d'Ivoire, Croatia, Cuba, Cyprus, Czech Republ, Czechoslovakia, Denmark, Dominican Re, Ecuador, Egypt, El Salvador, Estonia, Ethiopia, Federal Rep., Figi, Finland, Former USSR, Former Yugos, France, Gaza Strip, Georgia, Germany, Ghana, Greece, Guatemala, Guinea, Guinea-Bissa, Haiti, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Korea, North, Korea, outh, Laos, Latvia, Lebanon, Libya, Lithuania, Luxembourg, Madagascar, Malawi, Malaysia, Mali, Marocco, Mexico, Mozambique, Myanmar (Burm, Nepal, Netherlands, New Zealand, Niger, Nigeria, Norway, Pakistan, Paraguay, Peru, Philippines, Poland, Portugal, Romania, Russian ede, Rwanda, Sao Tome and, Saudi Arabia, Senegal, Slovak Repub, Slovenia, Somalia, South Africa, Spain, Sri Lanka, Suriname, Sweden, Switzerland, Syria, Tajikistan, Tanzania, Thailand, Total, Tunisia, Turkey, Uganda Ukraine United Kingd United State Uzbekistan Venezuela Vietnam, Yemen, Zaire, Zambia, Zimbabwe

III. List of variables, their definitions, sources and years available:

1. **Inflows of Foreign Population**
   Source: National statistical offices.
   Years available: 1985-2006
   Tim Hatton provided the extrapolated US migration data for years 1985-1989

2. **Stock of Foreign Population**

   Source: National statistical offices.

   Years available: 1985-2006

3. **GDP per capita, PPP (constant 2005 international $):** PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the U.S. dollar has in the United States. GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2005 international dollars.

   Source: World Bank, International Comparison Program database.

   Years available: 1984-2007

4. **Unemployment, total (% of total labor force):** Unemployment refers to the share of the labor force that is without work but available for and seeking employment. Definitions of labor force and unemployment differ by country.

   Source: International Labour Organisation, Key Indicators of the Labour Market database.

   Years available: 1984-2007

5. **Total population** is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship--except for refugees not permanently settled in the country of asylum, who are generally considered part of the population of their country of origin.

Source: World Bank staff estimates from various sources including census reports, the United Nations Statistics Division's Population and Vital Statistics Report, country statistical offices, and Demographic and Health Surveys from national sources and Macro International.

Years available: 1985-2006

6. **Public social expenditure as a percentage of GDP (SNA93):** Social expenditure is the provision by public institutions of benefits to, and financial contributions targeted at, households and individuals in order to provide support during circumstances which adversely affect their welfare, provided that the provision of the benefits and financial contributions constitutes neither a direct payment for a particular good or service nor an individual contract or transfer. Such benefits can be cash transfers, or can be the direct ("in-kind") provision of goods and services.

   Source: All data comes from the **OECD Social Expenditure Database (SOCX)**, with specific country notes also extracted from that database. More information is available under www.oecd.org/els/social/expenditure.

   Years available: 1985-2003

7. **Freedom House Index – Political Rights** represents scores of political rights and freedom. These are measured on a one-to-seven scale, with one representing the highest degree of freedom and seven the lowest.

   Source: Annual Freedom in the World Country Scores. Years 1985-2006

POLITICAL RIGHTS

Rating of 1 – Countries and territories with a rating of 1 enjoy a wide range of political rights, including free and fair elections. Candidates who are elected actually rule, political parties are competitive, the opposition plays an important role and enjoys real power, and minority groups have reasonable self-government or can participate in the government through informal consensus.

Rating of 2 – Countries and territories with a rating of 2 have slightly weaker political rights than those with a rating of 1 because of such factors as some political corruption, limits on the functioning of political parties and opposition groups, and foreign or military influence on politics.

Ratings of 3, 4, 5 – Countries and territories with a rating of 3, 4, or 5 include those that moderately protect almost all political rights to those that more strongly protect some political rights while less strongly protecting others. The same factors that undermine freedom in countries with a rating of 2 may also weaken political rights in those with a rating of 3, 4, or 5, but to an increasingly greater extent at each successive rating.

Rating of 6 – Countries and territories with a rating of 6 have very restricted political rights. They are ruled by one-party or military dictatorships, religious hierarchies, or autocrats. They may allow a few political rights, such as some representation or autonomy for minority groups, and a few are traditional monarchies that tolerate political discussion and accept public petitions.

Rating of 7 – Countries and territories with a rating of 7 have few or no political rights because of severe government oppression, sometimes in combination with civil war. They may also lack an authoritative and functioning central government and suffer from extreme violence or warlord rule that dominates political power.

Status of Free, Partly Free, Not Free – Each pair of political rights and civil liberties ratings is averaged to determine an overall status of "Free," "Partly Free," or "Not Free." Those whose ratings average 1.0 to 2.5 are considered Free, 3.0 to 5.0 Partly Free, and 5.5 to 7.0 Not Free (see table 3 in the "Checklist Questions and Guidelines" document). The designations of Free, Partly Free, and Not Free each cover a broad third of the available scores. Therefore, countries and territories within any one category, especially those at either end of the category, can have quite different human rights situations. In order to see the distinctions within each category, a country or territory's political rights and civil liberties ratings should be examined. For example, countries at the lowest end of the Free category (2 in political rights and 3 in civil liberties, or 3 in political rights and 2 in civil liberties) differ from those at the upper end of the Free group (1 for both political rights and civil liberties). Also, a designation of Free does not mean that a country enjoys perfect freedom or lacks serious problems, only that it enjoys comparably more freedom than Partly Free or Not Free (or some other Free) countries.

Years available: 1985-2006

8. **Freedom House Index – Civil Liberties** represents scores of civil liberties, and freedom. These are measured on a one-to-seven scale, with one representing the highest degree of freedom and seven the lowest.

Source: Annual Freedom in the World Country Scores. Years 1985-2006

CIVIL LIBERTIES

Rating of 1 – Countries and territories with a rating of 1 enjoy a wide range of civil liberties, including freedom of expression, assembly, association, education, and religion. They have an established and generally fair system of the rule of law (including an independent judiciary), allow free economic activity, and tend to strive for equality of opportunity for everyone, including women and minority groups.

Rating of 2 – Countries and territories with a rating of 2 have slightly weaker civil liberties than those with a rating of 1 because of such factors as some limits on media independence, restrictions on trade union activities, and discrimination against minority groups and women.

Ratings of 3, 4, 5 – Countries and territories with a rating of 3, 4, or 5 include those that moderately protect almost all civil liberties to those that more strongly protect some civil liberties while less strongly protecting others. The same factors that undermine freedom in countries with a rating of 2 may also weaken civil liberties in those with a rating of 3, 4, or 5, but to an increasingly greater extent at each successive rating.

Rating of 6 – Countries and territories with a rating of 6 have very restricted civil liberties. They strongly limit the rights of expression and association and frequently hold political prisoners. They may allow a few civil liberties, such as some religious and social freedoms, some highly restricted private business activity, and some open and free private discussion.

Rating of 7 – Countries and territories with a rating of 7 have few or no civil liberties. They allow virtually no freedom of expression or association, do not protect the rights of detainees and prisoners, and often control or dominate most economic activity.

Countries and territories generally have ratings in political rights and civil liberties that are within two ratings numbers of each other. For example, without a well-developed civil society, it is difficult, if not impossible, to have an atmosphere supportive of political rights. Consequently, there is no country in the survey with a rating of 6 or 7 for civil liberties and, at the same time, a rating of 1 or 2 for political rights.

Years available: 1985-2006

9. **Distance between countries –** capitals in km.

Source: MapInfo, own calculations.

10. **Neighbouring index –** in the form of dummy for neighbouring countries - value 1, 0 otherwise.

Source: MapInfo, own data gathering.

11. **Linguistic proximity:** the index for linguistic closeness between a pair of countries ranging from 0 to 1, depending on family of languages the two languages of destination and source country belong to. The index is constructed in the following way. First we define 4 weights:

- ***SAMEW1*= 0.1** if two languages are only related at the most aggregated first linguistic tree level, e.g. Indo-European versus Urallic languages (Finnish, Estonian, Hungarian);
- ***SAMEW2*= 0.15** if two languages belong to the second- linguistic tree level, e.g. Germanic versus Slavic languages;
- ***SAMEW3*=0.20** if two languages belong to third linguistic tree level, e.g. Germanic West vs. Germanic North.

- ***SAMEW4*=0.25** if two languages belong to fourth - highest level of language family, e.g. Scandinavian West (Icelandic) vs. Scandinavian East (Danish, Norwegian and Swedish), or German vs. English.

Further, we define linguistic index as :

INDEX= *SAMEW1* + *SAMEW2* + *SAMEW3* + *SAMEW4*

The index is equal to 0 if two languages do not belong to any common language family and equal to 1 for a common language in two countries. The linguistic index is based on information from Ethnologue: Languages of the World.

Source: Ethnologue: Languages of the World, 14th edition. http://www.ethnologue.com/web.asp, own data collection and calculation.

12. **Colony**– in the form of dummy for countries ever in colonial relationship – value 1, 0 otherwise.

Source: variable kindly provided by Andrew Rose, used for paper Rose, A. (2002): "Do We Really Know that the WTO Increases Trade?" NBER Working Paper No. 9273.