The role of language in shaping international migration: Evidence from OECD countries 1985-2006

Alicia Adsera

Princeton University and IZA

Mariola Pytlikova

Aarhus University, CCP and CIM

Abstract

In addition to economic determinants in line with neoclassical economics and the "human capital investment" framework, a number of non-economic factors are also relevant to explain migration decisions. Beside classic factors such as "love and wars", these include random events such as environment and climate shocks, migrant networks, language and aspects of "cultural distance". In that regard, the more "foreign" or distant the new culture is and the larger the language barrier is, the higher are the costs of an individual to migrate to a particular destination. Fluency in destination country's language and/or widely spoken languages (or ease to quickly learn it) plays a key role in the transfer of human capital from the source country to another country and boosts the immigrant's success at the destination's labor market. We use data on immigration flows and stocks of foreigners in 27 OECD destination countries from 130 source countries for the years 1985–2006. In addition to standard covariates from gravity models, we include a set of indices of language distance to study their association to the observed flows: (1) an index ranging from 0 to 1 that measures the distance between the family of languages of destination an source country; (2) the linguistic proximity measure proposed by Dyen between pairs of languages; (3) a dummy for destinations with a "widely spoken" language as the native language (4) indices on the number, diversity and polarization of languages spoken in both source and destination country, to proxy for the "potential" ease to learn a new language and of adaptation; (5) measures of the diversity of the existing stock and flows of migrants (weighted by languages).

JEL Classification: J61, F22, O15

Keywords: International migration, language.

1. INTRODUCTION

The determinants and consequences of migratory movements have been long discussed in the economic literature. Besides economic determinants in line with neoclassical economics and "the "human capital investment" theoretical framework, (Sjaastad, 1962; and further Harris and Todaro, 1970), a number of non-economic factors are also highly important regarding the migration decision. Beside classic factors such as "love and wars", these include random events, environment, climate, migrant networks, language and aspects of "cultural distance", see e.g Pedersen et al. (2008), Belot and Ederveen (2009) and Chiswick and Hatton (2003). Regarding the latter two factors, the more "foreign" or distant the new culture is and the larger the language barrier is, the higher are the costs of an individual to migrate and the less likely is it that the individual decides to migrate.

In particular, the ability to learn quickly and to speak a foreign language is an important factor in the potential migrants' decision making. Fluency in destination country language and/or widely spoken languages plays a key role in the transfer of human capital to a foreign country and generally it helps the immigrant to be successful at the destination country's labor market, see e.g. Kossoudji (1988), Bleakley and Chin (2004); Chiswick and Miller (2002, 2007), Dustmann (1994), Dustman and van Soest (2002) and Dustmann and Fabbri, (2003). By exploiting differences between younger and older arrivers as effects of language skills, Bleakley and Chin (2004 and 2010) find that language knowledge is a key for outcomes of immigrants in terms of education, earnings and social outcomes. A study by Adsera and Chiswick (2007) found that there is around 9 per cent earnings premium for immigrant men if they come from a country where the language spoken belongs to the same language family group as the destination country. Thus the linguistic skills and linguistic proximity seem to be very important in accounting for migrants well-being. However, previous evidence on the determinants of migration hardly ever went beyond the inclusion of a simple dummy for sharing a common language.

The main contribution of this paper is to investigate the role of linguistic proximity in migrant's decision making by using a wide range of linguistic indicators. Among others we make use a more refined indicator of the linguistic distance between two countries based on the family of languages to which both the official and any other spoken languages belong to. Further we make use of the linguistic proximity measure proposed by Dyen et al. (1992), a group of linguists who built a measure of distance between Indo-European languages based on the proximity between samples of words from each language. Finally, we control for the fact that potential migrants prefer to choose a destination with a "widely spoken" language as the native language. The rationale behind the last mentioned variable is the following: first, a knowledge of particular foreign languages increases chances of a potential immigrant to be successful at the foreign labour market and helps to lower his/her costs of migration, as discussed above. Second, foreign language proficiency might be considered an important part of human capital in the labour market of the source country. Thus, the learning/practicing/improving of a "widely spoken" language in the "native" countries serves as a pull factor especially for temporary migrants. 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.

The rest of the paper is organized as follows: Section 2 surveys earlier research in the area and 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 collected for this study and the independent variables included in the analysis. Results from the econometric analyses are given in Section 5. Finally, Section 6 offers some concluding remarks.

2. THEORY AND PREVIOUS RESEARCH ON MIGRATION DETERMINANTS

The determinants and consequences of migratory movements have been long discussed in the economic literature. The first contributions can be found in the neoclassical economics, which stress differentials in wages as a primary determinant of migration (Hicks, 1932). The "human capital investment" theoretical framework (Sjaastad, 1962) adds the existence of migration costs to the migrants' decision making model, so that a person decides to move to another country only if the discounted expected future benefit is higher than the costs of migration. The "human capital investment" theoretical framework has been further adjusted for the probability of being employed; see Harris and Todaro (1970). In aggregate terms, the differentials in wages and probability of being unemployed are typically proxied by GDP per capita levels in destination and source countries and unemployment rates[1], respectively. The effect of GDP per capita in the source country may be more mixed. Earlier studies have found an inverted 'U' relationship between source country

[1] Sometimes employment or vacancy rates are used instead of unemployment rates.

GDP and emigration, see Hatton and Williamson (2005) and Pedersen et al. (2008). At very low levels of GDP, emigration is low because people are too poor to pay the migration costs. At higher income levels, migration increases, and as GDP levels increase further, migration may again decrease because the economic incentives to migrate to other countries decline.

In line with the human capital framework, empirical studies confirm that socio-demographic characteristics of an individual such as age, gender and education[2] matter in the decision to migrate. Usually, the young and more educated individuals are more mobile – as they have higher "returns to migration". Thus, the socio-demographic structure of a source country population matters, see e.g. Chiswick (2000), Fertig and Schmidt (2000), Bauer and Zimmerman (1999) and Krieger (2004).

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 is that potential emigrants must take into account the probability of being unemployed in the destination country. The damaging consequences of this risk 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), Urrutia (2001), Pedersen et al. (2008), among others. Besides, immigration policies and changes in the policies over time strongly contribute to shape migration flows as they differ between potential receiving countries (Mayda, 2010).

The costs of migration are also shown to be an important part of migrants' decision making. The migration costs are not only the out-of-pocket expenses, but also psychological costs connected to moving to a foreign country and leaving family, friends and the known environment. The costs typically increase with the physical distance between two countries. However, changes and improvements in communication technologies and declining transportation costs may imply that the effect of "distance" has been reduced during the latest decades. Further, network effects may also counteract "distance". Through "networks" potential migrants receive information about the immigration country - about the possibility of getting a job, economic and social systems, immigration policy, people and culture. This facilitates immigration and the adaptation of new immigrants into the new environment. Network effects may also help to explain the persistence of migration flows, see e.g. Epstein (2002), Bauer et al. (2002) and Heitmueller (2003). Empirical evidence has shown that migrant networks have a significant impact on sequential migration, see

[2] It is argued that a more educated individual has a greater ability to e.g. collect and process information, which lowers the risk and increases the propensity to migration, or to learn a foreign language.

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.

In addition to that, the linguistic and cultural distance is important as well. The more “foreign” or distant the new culture and the larger the language barrier is, the higher are the costs of an individual to migrate and the less likely is it that the individual decides to migrate, holding all other factors constant (Pedersen et al., 2008). A recent study by Belot and Ederveen (2010) show that cultural barriers 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 factor in the potential migrants’ decision making. Fluency in destination country language and/or widely spoken languages plays a key role in the transfer of human capital to a foreign country and generally it helps the immigrant to be successful at the destination country’s labor market, see e.g. Kossoudji (1988), Bleakley and Chin (2004); Chiswick and Miller (2002, 2007), Dustmann (1994), Dustman and van Soest (2002) and Dustmann and Fabbri, (2003). By exploiting differences between younger and older arrivers as effects of language skills, Bleakley and Chin (2004 and 2010) find that language knowledge is a key for outcomes of immigrants in terms of education, earnings and social outcomes. Study by Adsera and Chiswick (2007) found that there is around 9 per cent earnings premium for immigrant men if they come from a country, where the language spoken belongs to the same language family group as the destination country. Thus the linguistic skills and linguistic proximity seem to be very important. Besides, destination countries with a “widely-spoken” language of natives can act as pulls in international migration. There may be two different forces behind the migration pattern. As some of the “widely spoken” languages are often taught at schools in many source countries, the immigrants are more likely to migrate to destinations, where the languages are spoken. Second, the foreign language proficiency is considered to be an important part of human capital at 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. Thus, the learning/practicing/improving the skills of “widely spoken” language in the “native” countries serve as a pull factor especially for temporary migrants.

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 capacity to absorb further more languages (Kovacs and Mehler, 2009).). If this is the case we should expect, 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 this paper), other things being constant.

At the same time an increase diversity of language at origin may also be a proxy for ethnic or political fractionalization that can by itself be 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 controversial over this issue i.e. 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 should be related to whether political tension may be associated or not to 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 pure measures of inequality or diversity (ie, income, ethnic groups...) to understand political demand and civil strives...etc. Similarly Desmet et al. (2009 a & b) 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 we use both measures of diversity and of polarization developed by Desmet et al. (2009) that take into account linguistic distances across the different groups in a society to understand whether both forces may be at play. It might be that larger linguistic polarization correlated with more conflicts, lower trust measures and lower economic growth, can have consequently a negative effect on migration. It has been shown in previous studies, e.g. Hatton and Williamson (2005) and Pedersen et al.(2008), that source country GDP has an inverted U-shape effect on migration due to poverty constrains to cover costs of migration[3].

Similarly, the diversity and polarization of languages at the destination country may make it more or less attractive to the potential migrant. Again, a largely polarized society may increase the costs of adaptation, once linguistic distance of the migrant is taken into account. But diversity per se, if the linguistic distance of the different groups is not large, should not pose the same problem. A

[3] At higher income levels, migration increases, and when GDP levels increase further, migration may again decrease because the economic incentives to migrate to other countries decline.

diverse society might have in place more flexible policies that adapt to the needs of different constituencies (i.e., education immersion in different languages according to the area of the country to facilitate adaptation of newcomers).

Finally, the composition and diversity of the migrants already present in a given destination may affect the likelihood that a potential migrant finds previous migrants from his/her same country and/or linguistic groups. A larger community of people of their own linguistic background facilitates the initial entry into the labor market. Many immigrants may even spend their whole lives working in a linguistic enclave within their destination location (i.e. Boyd 2010 for the case of Canada). Also a more diverse destination may be more ready to receive a newcomer and his/her family with regard to public services, language training and children's education. In addition, if the linguistic community of a migrant at destination is large (even if existing migrants are not coming from the same country), networks and linguistic enclaves may facilitate labor market entry to newcomers (i.e. migrants for all Central America moving to highly Mexican areas in the US).

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 exploring the different dimensions of the link.

3. A MODEL OF INTERNATIONAL MIGRATION

A 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) 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*

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

where S_{ijkt} is a vector of characteristics that affects an individual's 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 or family members are. D_{ij} reflects time-independent fixed-out-of-pocket and psychological/social costs of moving from country *i* to country *j*. X_{ikt} and X_{jkt} are vectors of push and pull factors that vary across time and affect individual *k's* choice where *i* denotes source country and *j* denotes destination country, (*i* = 1,…,130, and *j* = 1,….,25); *t* is time period (*t* = 1,…,22). We assume the utility of an individual has a linear form:

$$U_{ijkt} = \alpha_1 S_{ijkt} + \alpha_2 D_{ij} + \alpha_3 X_{ikt} + \alpha_4 X_{jkt} + \varepsilon_{ijkt} \qquad (2)$$

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. 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 25 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 for estimation of 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,* i.e. 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:

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

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

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

4. 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, see Pedersen, Pytlikova and Smith (2008) for a detailed description of the dataset[4].

Besides the flow and stock information, the dataset contains a number of other time-series variables, which might help to explain the migration flows between the countries. These variables were collected from different sources, e.g. OECD, the World Bank and others; see Appendix for definitions, sources of the variables and summary statistics. For a more comprehensive description of the dataset, see Pedersen et al. (2008).

Departing from equation (5), we normalize the immigration flows by population size in source country, i.e. we use the emigration rate, m_{ijt}, instead of migration flow in absolute numbers as the dependent variable. All time-varying explanatory variables are lagged by one year in order to account for information, on which the potential immigrants base their decision to move. In this way, we also reduce a risk of simultaneity bias.

Further, we include the normalized lagged stock of immigrants, S_{ijt-1}, i.e. the stock of immigrants from source country *i*, divided by population in source country *I,* as a proxy for "networks".

Finally, in most of the models we include a full set of destination and source fixed effects, $c_{j,}$ and $c_{i,}$ in order to capture unobserved time constant factors influencing immigration flows, for instance differences in national immigration policy. Thus, the model to be estimated is:

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

D_{ij} contains variables reflecting costs of moving to a foreign country. First, we include a variable describing cultural similarity denoted *Neighbourg Country.* It is a dummy variable assuming the value of 1 if the two countries are neighbours, 0 otherwise. The variable *Colony* is a dummy variable assuming the value of 1 for countries ever in colonial relationship, 0 otherwise. This variable is included because the past colonial ties might have some influence on cultural distance: provide better information and knowledge of potential destination country and thus lower migration

[4] The original OECD migration dataset by Pedersen, Pytlikova and Smith (2008) covers 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 purposes of the paper we additionally included Slovenia as country of origin and 5 other OECD countries as destinations – Czech and Slovak Republics, Hungary, Poland and Ireland. Further, we extended the existing time period by the years 1985-1989 and 2001-2006.

costs, which could encourage migration flows between these countries. In order to control for the direct costs (transportation costs) of migration we use the measure of the *Distance in Kilometres* between the capital areas in the sending and receiving countries.

The explanatory variables included in the X_{it-1} and X_{jt-1} cover a number of push and pull factors such as the economic development measured by GDP per capita in destination and source countries (which are supposed to capture the relative income opportunities in the two countries), employment opportunities in the sending and receiving countries, measured by unemployment rates, and relative size of populations in destination and source countries.. Additionally as a pull factor we include information on the extent of welfare provisions in the country of destination measured as public social expenditure as percentage of GDP.

The political pressure in the source country may also influence migration. Therefore, we include a couple of index variables from *Freedom House* which intend to measure first, the degree of freedom, 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 flows out of a given country.

Our linguistic variables of interests are covered in matrix *L.* We include a variable *Linguistic Distance,* which is an index ranging from 0 to 1, depending on family of languages the two languages of destination and source country belong to. The index is equal to 0 if two languages do not belong to any common language family and to 0.1 if they only are related at the most aggregated linguistic level (ie. Indo-European versus Dravidian). Further the weights for creating an index are: 0.25, 0.45 and 0.7 if the languages both belong to the second-, third- and fourth - highest level of language family, respectively. We set the index equal to 1 for a common language in two countries. The linguistic index is based on information from Ethnologue, and will be described in detail in Appendix section.

Many countries have more than one official language and among those one is the most widely used. To construct the index we use the language most extensively used in the country. As a part of robustness analyses, we extend the set of linguistic measures to include an index that takes into account the existence of multiple official languages and set 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 according the most widely used language in that 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 more proxy to their own and that enhances their labor market returns. In addition, we also make use of linguistic proximity measure proposed by Dyen et al. (1992), a group of linguists who built a measure of distance between Indo-European languages based on the proximity between samples of words from each language. 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 in specifications containing this variable is severely reduced since only countries with Indo-European languages are included.

Further, in the regressions presented here we control for the fact that potential migrants prefer to choose a destination with a “widely spoken” language as the native language and we include dummies for widely spoken language (in particular, English, Spanish, German, French, Italian and Portuguese).

To account for the diversity of indices in both the country of origin and destination we use a couple of indices from Desmet et al (2009b) to measure diversity: fractionalization and polarization. Desmet et al. (2009 b) use linguistic trees, describing the 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. 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, fractionalization is just the probability that two individuals chosen at random, will belong to different groups.

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

This measure is maximized when each individual belongs to a different group. Polarization, in contrast, is maximized when there are two groups of equal size. So if a country A consist 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. We use the polarization measure in Desmet at al. (2009b) that is derived from Montalvo and Reynal-Querol (2005). This index satisfies the conditions for a desirable index of polarization in the axiomatic approach of Esteban and Ray (1994)

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

Even though Desmet et. al (2009b) calculate these indices for 15 different levels of aggregation, in the paper we only use two of their measures at the 1st and 4th levels of aggregation of linguistic families available in the linguistic classification of Ethnologue. For space limitations, tables in section 5 present results only at level 4.[5] In addition we use two more measures from Desmet er al. (2009a), GI diversity and ER polarization indexes, which allow controlling for distances between different linguistic groups. The GI index was proposed by Greenberg (1956) 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 groups. Desmet et al (2009a) define the distances by the number of potential linguistic branches that are shared between the languages of two groups. Similarly the 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.

Finally, we add measures of the number of languages, in order to account for information on intensity of multilingualism in a given country We use two different measures: the number indigenous languages obtained from Ethnologue, and another index, which limits the number of languages at the linguistic tree level 2 to those spoken by a given minimum 5 per cent of a countries population. [6]

[5] The implied diversity of the index changes somewhat as the level of linguistic aggregation varies. Desmet et al. (2009b) 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)”.

[6]The measures on number of language at different linguistic levels, spoken by different percentage of population were given to us by Ignacio Ortuno-Ortin.

Finally to account for the diversity of the stock and flows of migrants to a destination country we calculate a set of time-varying Herfindahl-Hirschman indices for both measures by country (HI and HI flows, respectively). We have also calculated similar diversity indices weighted by language of migrant groups. In those indices we group together all migrants with similar linguistic background regardless of their country of origin. In addition we introduce a measure of migrant linguistic community, defined as the stock of all migrants in a country with a similar language than the newcomer, regardless of their country of origin. For these last two measures we have experimented with different levels of linguistic differences, at levels 3 and 4, in particular. Estimates presents results with the stock of migrants of the same linguistic group at a branch of level 4 and also measures of linguistic diversity of migrants at the same level.

5. RESULTS

Econonometric specification

Table 1, columns 1 to 5, shows pooled OLS estimates of different model specifications from parsimonious to full specification excluding unemployment rates[7]. All specifications contain a time trend variable[8] and have "robust" Hubert/White/sandwich standard errors clustered at each pair of destination and source country.

Our variable of interest, the linguistic distance variable, attaches a significant positive coefficient across all specifications[9]. Thus, other things being equal, emigration flows between two countries are larger the closer their linguistic distance. 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 distance decreases to only around 0.14, but continues to be highly significant. Another reason for which two countries can be relatively closely related is sharing a colonial past or being geographically close. Also, some former colonies may have adopted the language of their colonial power. In column (3) we add dummies for past colonial relationship

[7] The reason why we would like to show the results without the unemployment variables is that the source country unemployment rates imposes the largest restriction with respect to the number of missing observations. By excluding unemployment variables we gain double of the size compared to the full model specification.

[8] In some other tables we added year dummy and trend variables in order to control for common idiosyncratic shocks over the time period that we analyze. The dummies didn't add much to the results; therefore we do not report the results, they are available from the authors upon request.

[9] Alone the linguistic distance explains approximately 9 % of the explanatory power (adj. R-squared).

between both countries as well as measures of distance between capitals and an indicator of whether they share common borders. Among these variables, only distance and colonial past is clearly significantly associated with weaker and stronger emigration flows, respectively. The coefficient of linguistic distance is only slightly affected by their inclusion, see column (3). Part of the influx of new migrants into a country may be driven 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. The coefficient to the linguistic distance drops to 0.03 when including the lagged stock of immigrants variable. Accounting for recent flows of immigrants to the country (lagged value of flows) in column (5) also reduces the coefficient further to 0.007 but it continues to be highly significant at less than 0.001.

Besides the variables considered in our full model, there are other unobservable factors that shape international migration flows and that are characteristic for particular countries. To account for the unobserved country-specific heterogeneity, we add destination and origin country fixed-effects. In the context of international migration, there is a question whether to account for destination- and origin-country specific effects, μ_i and μ_i, or pair of countries specific effects, μ_{ij}. Destination and origin country unobservable effects might represent characteristics of immigration policy practices in each destination country, as well as climate, weather, openness towards foreigners or culture. On the other hand, pair of countries unobservable effect might capture traditions, historical, and cultural ties between two particular countries of destination and origin, as well as bilateral immigration policy schemes between the two countries. However, as our main focus in the paper is on linguistic and cultural distance and their effect on migration, and the unobserved pair of countries specific fixed effects would be picking up the effect, our preferred specification is the model with destination and origin country fixed effects with clustered standard errors on the level of pair of countries[10].

[10] This is our preferable specification also from the statistical point of view: besides taking the effects from variables of interest pair of countries effects mean too many parameters.

In Table 2 we present results of our full model specification – also with unemployment rates[11]. In the table we analyze the stability of the results with respect to the choice of different econometric specifications. In the first three columns we show OLS estimates. In columns 4 and 5 we present estimates of non linear least squares.[12] And, finally, in column 6 and 7 we include destination and source country fixed effects into the OLS and NLS estimations, respectively. When comparing the pooled OLS results with respect to linguistic diversity with the NLS results and the panel models treating destination and source countries as fixed effects, the overall impression is that the results regarding sign and statistical significance are quite robust across the different specifications. However, the absolute sizes of the coefficients to the linguistic diversity are generally much larger when applying NLS. Also the panel data estimators which control for destination and source country-specific effects are larger in numerical magnitude.

As a part of robustness analyses, we extend the set of linguistic distance measures to include an index that takes into account the existence of multiple official languages and set the index at 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 measure of distance between Indo-European languages based on the proximity between samples of words from each language. However, the Dyen index covers only Indo-European languages and thus our number of observations is reduced significantly. The results are shown in Appendix table , with columns (1)-(3) containing results of OLS, FE and NLS regressions with multiple languages index and columns, respectively and columns (4)-(6) regressions with the Dyen index. The coefficient of the linguistic distance when using the multilanguage criteria is significant and positive only in the fixed effect regression. In OLS and NLS the coefficients to linguistic distance are statistically insignificant. On the other hand the Dyen index attaches a significant positive coefficient across all econometric specifications in columns (4) to (6) and its magnitude is in fact much larger then coefficients to our preferable linguistic distance measure as shown in tables 1 and 2. There are two potential explanations for the particular strength of this result. First, the sample is reduced to likely more homogeneous countries, since it excludes those with non-Indo-European languages. Second, the Dyen measure allows for greater variance across country-pairs

[11] Note that the number of observations decreases from approximately 27 thousands to 16 thousand due to missing observations for source country unemployment rates.

[12] Given the nonnegative nature of the data and its non-normal distribution across the sample, we estimate a NL specification where the level of migration flows is explained by the exponential of the linear combination of all log-transformed independent variables.

since it measures more continuously the distance between languages than the other measures in the paper.

Table 4 includes dummies for the major languages in destination countries to test for the hypothesis of “widely used” language. The table presents OLS estimates in columns (1) to (3) and NL estimates in column (4). In the most basic specification of column (1), without any other language indices or controls for existing stock or lagged dependent variable, all destinations with major languages, with the exception of Portuguese, have above average migration. The sign reverts for French once the linguistic distance between destination and source country is included in column (2). Interestingly, once all the other push and pull factors are included only English, German and Spanish generate a greater attraction than others. If anything, Portuguese and French seem to be associated with under average flows ceteris paribus. The sign also reverses for German in the nonlinear estimates of column (4). Overall, the prevalence of English and Spanish at a destination is regarded as an additional amenity.

Table 5 includes some of the measures of the diversity of the stock of migrants that we have explored. Each one of the measures is presented with four specifications, OLS and NL with and without country of destination and origin fixed effects. Columns (1) to (4) include an index of the diversity of the stock of migrants into the destination country measured with the log of a Hinferdahl-Hirschman index. In column (1) the positive coefficient implies that in places where the flow of migrants is more diverse, flows are smaller. However, once fixed effects are included in column 2 and also in the nonlinear specifications of columns (3) and (4), the coefficient on the diversity of the stock of migrants is negative, indicating that more migrants are flowing to places with a more diverse stock of previous movers, ceteris paribus. This seems to indicate that migrants may move to places that have better suited policies to attract movers from very diverse backgrounds. Nonetheless, the finding may also be in part endogenous since larger flows from every source country, ceteris paribus, should produce more diverse stock of migrants. Columns (5) to (8) present similar models that employ the diversity of the stock of migrants partitioned not by country of origin but by the linguistic tree at level 4. Thus migrants from different countries but whose language belongs to the same linguistic family at level 4 are counted as part of the same group. Again, the coefficient is positive in the basic OLS but turns to negative in the other specification indicating larger migration flows to countries with a linguistically more diverse stock of migrants. However the coefficient is only significant in the NL model without fixed effects.

Finally, columns (9) to (12) include the share of migrants in the country of destination whose language belongs to the same linguistic family (at level 4) than that spoken by the migrant. Even if the coefficient is not significant in the specifications without fixed effects, it is highly significant and positive when both origin and destination country dummies are included. Not surprisingly, individuals are choosing to move to destinations where a large share of the migrants speaks a language relatively close to theirs. The calculation of such a choice is that moving cost are lower and adaptation easier with an extensive local community that is culturally close. Moreover, it provides opportunities for working in "ethnic enclaves", particularly for those who envision their move as a temporary one. Additional results where we employ either the diversity of recent migrant flows or the diversity of the stock of migrants grouped at other levels of the linguistic tree are available upon request.

Table 6 includes a set of measures of the linguistic diversity 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 distance, network, economic conditions, distance variables and a time trend. Each model is first estimated by OLS and then with the nonlinear specification. None of the models includes fixed effects. 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 diversity of languages at the level 4 of the linguistic tree and is obtained from Desmet (2009b). The following two columns use the GI index from Desmet (2009 a) instead. That index takes into account the actual distance of languages and not only the particular linguistic family to which they belong as the Elf indices. In any case results are very similar for both measures. 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 the 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 diversity as an asset that facilitates language acquisition at destination and lowers migration costs.

The second row in Table 6 includes two sets of polarization indices, both at destination and at origin. The first index measures polarization at the 4th level in the linguistic tree of Ethnologue and the second measure ER (of the family of polarization measures started by Esteban and Ray) 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. Results are fairly similar to those for diversity indices. A more polarized linguistic environment at destination seems to deter migration flows, other things being the same.[13] Conversely, more polarized societies seem to significantly push larger number of people in the search of a new life elsewhere. Only the coefficient for the NL specification with the ER index reverses its sign and we have no good explanation for this result.

Finally two additional variables measuring the linguistic richness both country of origin and destination are presented in the third row of Table 6. Both the total number of indigenous languages or the number of languages at the linguistic three level 2 spoken by at least 5 per cent 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 from positive in OLS to negative in NL models.

6. CONCLUSIONS and FURTHER STEPS

(to be added)

[13] Results for polarization are somewhat stronger when polarization is measured at higher levels of the linguistic tree since the distance between those groups is larger than at the 4th level. Results are available upon request.

<u>References:</u>

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.

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

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.*

Boyd () Linguistic enclaves in Canada

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

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” (with Paul W. Miller), *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,” (with Paul W. Miller), *Review of the Economics of the Household*, 5 (2), June 2007, pp. 129-157.

Desmet, K., I. Ortuño-Ortín and R.Wacziarg (2009a), The Political Economy of Ethnolinguistic Cleavages, *NBER Working Paper* 15360.

Desmet, K., I. Ortuño Ortín and S. Weber (2009b), "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 (2002): “Language and the Earnings of Immigrants.” *Journal 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 Indoeuropean 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 Confict,” mimeo.

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

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

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.

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*. 6(3):205-228.

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

Mayda A.M. (2010): “International migration: A panel data analysis of the determinants of bilateral flows”, forthcoming in the *Journal of Population Economics*.

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.

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.

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

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

Table 1: OLS and FE (destinations and origins) Estimation of migration flows from 130 countries of origin(i) to 25 OECD destination countries(j), 1985-2006.

	OLS	OLS	OLS	OLS	OLS	FE
	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	EFlows	EFlows	EFlows	EFlows	EFlows	EFlows
lnindex2	0.480***	0.137***	0.097***	0.031***	0.007***	0.019***
	(0.030)	(0.024)	(0.023)	(0.012)	(0.003)	(0.006)
EFlows_1					0.848***	0.777***
					(0.007)	(0.009)
EStock_1				0.719***	0.096***	0.131***
				(0.009)	(0.006)	(0.007)
GDPpCapPPPj1		2.089***	2.130***	0.493***	0.129***	1.218***
		(0.101)	(0.097)	(0.059)	(0.015)	(0.088)
GDPpCapPPPi1		0.480***	0.535***	-0.002	0.009**	0.009
		(0.028)	(0.037)	(0.020)	(0.004)	(0.037)
lnpsepj1		-0.672***	-0.651***	-0.154**	-0.069***	-0.041
		(0.139)	(0.134)	(0.074)	(0.014)	(0.051)
lnEPop_ij1		0.629***	0.629***	0.130***	0.025***	-0.108*
		(0.016)	(0.016)	(0.010)	(0.002)	(0.058)
distance		-0.461***	-0.331***	-0.234***	-0.053***	-0.108***
		(0.035)	(0.037)	(0.021)	(0.005)	(0.010)
neighbour			1.346***	0.019	-0.001	-0.024
			(0.156)	(0.086)	(0.017)	(0.019)
colony			2.034***	0.186	0.112***	0.099***
			(0.176)	(0.129)	(0.030)	(0.031)
lnFreedomPRi1			-0.139**	0.001	-0.019*	0.006
			(0.068)	(0.040)	(0.010)	(0.013)
lnFreedomCRi1			0.287***	0.001	0.033***	0.001
			(0.081)	(0.049)	(0.012)	(0.016)
Destination & Origin FE	NO	NO	NO	NO	NO	YES
year	0.020***	-0.027***	-0.027***	-0.005***	-0.001	-0.024***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.001)	(0.002)
Constant	-43.880***	24.805***	22.564***	4.321	0.205	36.750***
	(5.376)	(5.163)	(5.103)	(3.732)	(1.056)	(3.666)
Observations	45458	39737	39313	26822	25651	25651
Adjusted R-squared	0.086	0.540	0.573	0.877	0.960	0.962

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

Table 2: OLS, NLS and FE(destinations and origins), adding unemployment rates. Estimation of migration flows from 130 countries of origin (i) to 25 OECD destination countries (j), 1985-2006.

	OLS	OLS	OLS	NLS	NLS	FE	NLS FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	EFlows	EFlows	EFlows	EMjit	EMjit	EFlows	EMjit
lnindex2	0.086***	0.025*	0.006*	0.157***	0.0385**	0.017***	.0750*
	(0.028)	(0.015)	(0.003)	(0.0580)	(0.0157)	(0.006)	(0.041)
EFlows_1			0.864***		0.835***	0.794***	.7438
			(0.009)		(0.0304)	(0.012)	(0.050)
EStock_1		0.733***	0.086***	0.701***	0.0674***	0.128***	.0094
		(0.012)	(0.007)	(0.0495)	(0.0224)	(0.009)	(0.068)
GDPpCapPPPj1	2.513***	0.277***	0.135***	-1.383***	-0.325**	1.226***	.5676
	(0.133)	(0.084)	(0.019)	(0.389)	(0.153)	(0.122)	(0.262)
GDPpCapPPPi1	0.307***	-0.094***	-0.020***	-0.585***	-0.263***	-0.056	-.7260
	(0.057)	(0.029)	(0.007)	(0.143)	(0.0851)	(0.040)	(0.186)
lnpsepj1	-1.052***	0.058	-0.077***	0.422	-0.0900	-0.113*	-.8245
	(0.170)	(0.098)	(0.018)	(0.464)	(0.108)	(0.060)	(0.269)
UnemplRate_j1	0.518***	-0.105***	0.045***	0.0198	0.0164	0.052***	.0839
	(0.069)	(0.038)	(0.010)	(0.155)	(0.0747)	(0.014)	(0.104)
UnemplRate_i1	0.117**	0.028	0.004	0.0386	-0.0767*	0.032**	-.1591
	(0.053)	(0.028)	(0.006)	(0.123)	(0.0403)	(0.014)	(0.077)
lnEPop_ij1	0.578***	0.116***	0.016***	0.0682	0.0251	-0.115	.0361
	(0.019)	(0.012)	(0.003)	(0.0456)	(0.0169)	(0.084)	(0.058)
distance	-0.341***	-0.190***	-0.044***	0.0206	-0.0382	-0.087***	-.1003
	(0.040)	(0.023)	(0.005)	(0.113)	(0.0251)	(0.010)	(0.052)
neighbour	1.358***	0.068	0.003	0.325	0.00256	-0.028	.01967
	(0.157)	(0.088)	(0.016)	(0.203)	(0.0620)	(0.018)	(0.091)
colony	1.688***	0.218	0.100***	-0.907**	0.0720	0.074**	.3557
	(0.212)	(0.147)	(0.032)	(0.405)	(0.0717)	(0.033)	(0.169)
lnFreedomPRi1	0.025	-0.027	-0.027**	-0.105	-0.260***	0.001	-.3921
	(0.081)	(0.047)	(0.011)	(0.145)	(0.0948)	(0.018)	(0.113)
lnFreedomCRi1	0.012	-0.085	0.006	-0.188	-0.0950	-0.018	.0462
	(0.091)	(0.053)	(0.013)	(0.171)	(0.0813)	(0.019)	(0.095)
Destination and Origin FE	NO	NO	NO	NO	NO	YES	YES
trend	-0.035***	0.011***	0.003***	0.00801***	0.00320***	-0.021***	.00195
	(0.004)	(0.003)	(0.001)	(0.00231)	(0.00111)	(0.003)	(0.0015)
Constant	37.160***	-25.953***	-6.410***			31.253***	
	(6.815)	(4.822)	(1.333)			(4.462)	
Observations	23102	16814	16221	16814	16221	16221	16221
Adjusted R-squared	0.572	0.872	0.963	0.679	0.908	0.965	0.918

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

Table 3: Robustness checks – using alternative indexes for linguistic distance: IndexAll and Dyen index. Estimation of migration flows from 130 countries of origin (i) to 25 OECD destination countries (j), 1985-2006. OLS and NLS.

	OLS	NLS	FE	OLS	NLS	FE
	(2)	(4)	(6)	(2)	(4)	(6)
VARIABLES	EFlows	EMjit	EFlows	EFlows	EMjit	EFlows
lnindexAll	-0.004	0.00302	0.017**			
	(0.003)	(0.0160)	(0.007)			
lndyen				0.035***	0.0898**	0.041***
				(0.009)	(0.0364)	(0.012)
EFlows_1	0.864***	0.838***	0.795***	0.905***	0.810***	0.858***
	(0.009)	(0.0309)	(0.012)	(0.008)	(0.0327)	(0.012)
EStock_1	0.086***	0.0810***	0.127***	0.054***	0.0878**	0.083***
	(0.007)	(0.0231)	(0.010)	(0.007)	(0.0357)	(0.010)
GDPpCapPPPj1	0.143***	-0.379**	1.227***	0.145***	-0.341	1.158***
	(0.019)	(0.159)	(0.123)	(0.024)	(0.212)	(0.176)
GDPpCapPPPi1	-0.021***	-0.231***	-0.058	-0.020*	-0.374***	-0.042
	(0.007)	(0.0803)	(0.040)	(0.010)	(0.121)	(0.068)
lnpsepj1	-0.082***	-0.121	-0.113*	-0.066***	-0.0627	-0.170**
	(0.018)	(0.118)	(0.060)	(0.021)	(0.131)	(0.079)
UnemplRate_j1	0.046***	0.0212	0.052***	0.084***	0.00620	0.094***
	(0.010)	(0.0742)	(0.014)	(0.012)	(0.0979)	(0.018)
UnemplRate_i1	0.007	-0.0496	0.032**	0.029***	-0.0537	0.039**
	(0.006)	(0.0358)	(0.014)	(0.009)	(0.0616)	(0.018)
lnEPop_ij1	0.016***	0.0325*	-0.113	0.007**	0.00164	-0.087
	(0.003)	(0.0169)	(0.084)	(0.003)	(0.0275)	(0.111)
distance	-0.043***	-0.0330	-0.084***	-0.047***	-0.000809	-0.073***
	(0.005)	(0.0243)	(0.010)	(0.005)	(0.0311)	(0.010)
neighbour	0.011	0.0132	-0.028	-0.036**	0.0605	-0.047***
	(0.016)	(0.0615)	(0.019)	(0.017)	(0.102)	(0.018)
colony	0.112***	0.109	0.076**	0.091***	-0.0286	0.035
	(0.033)	(0.0696)	(0.033)	(0.031)	(0.0712)	(0.030)
lnFreedomPRi1	-0.030***	-0.258***	0.001	-0.026*	-0.340**	-0.042*
	(0.012)	(0.0963)	(0.018)	(0.016)	(0.134)	(0.022)
lnFreedomCRi1	-0.003	-0.118	-0.018	0.016	-0.0850	0.006
	(0.013)	(0.0826)	(0.019)	(0.015)	(0.0982)	(0.021)
Destination and Origin FE	NO	NO	YES	NO	NO	YES
trend	0.002***	0.00324***	-0.021***	0.003***	0.00338**	-0.017***
	(0.001)	(0.00110)	(0.003)	(0.001)	(0.00158)	(0.004)
Constant	-5.932***		31.013***	-7.706***		24.078***
	(1.339)		(4.474)	(1.631)		(5.484)
Observations	16221	16221	16221	8900	8900	8900
Adjusted R-squared	0.963	0.907	0.965	0.967	0.904	0.969

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

Table 4: Widely spoken languages. Estimation of migration flows from 130 countries of origin (i) to 25 OECD destination countries(j), 1985-2006. OLS and NLS.

	OLS	OLS	OLS	NLS
	(1)	(2)	(3)	(4)
VARIABLES	EFlows	EFlows	EFlows	EMjit
lnindex2		0.113***	0.004	0.0310**
		(0.027)	(0.003)	(0.0127)
Engl	1.916***	1.981***	0.043**	-0.0941
	(0.122)	(0.142)	(0.022)	(0.0740)
Span	2.318***	1.108***	0.193***	0.0625
	(0.218)	(0.198)	(0.030)	(0.119)
Ger	0.862***	0.510***	0.046***	-0.101*
	(0.156)	(0.109)	(0.012)	(0.0536)
French	0.209*	-0.158*	-0.003	-0.127*
	(0.125)	(0.094)	(0.011)	(0.0653)
Ital	1.027***	0.655***	0.020	0.0792
	(0.150)	(0.121)	(0.014)	(0.0676)
Port	-0.514**	-0.622***	-0.095***	-1.019***
	(0.229)	(0.144)	(0.030)	(0.352)
EFlows_1			0.859***	0.833***
			(0.009)	(0.0303)
EStock_1			0.085***	0.0747***
			(0.008)	(0.0246)
GDPpCapPPPj1		1.736***	0.092***	0.234***
		(0.144)	(0.020)	(0.0785)
GDPpCapPPPi1		0.234***	-0.019***	-0.223***
		(0.039)	(0.007)	(0.0862)
lnpsepj1		0.714***	-0.029	-0.0618
		(0.222)	(0.025)	(0.191)
UnemplRate_j1		0.218***	0.011	0.135*
		(0.074)	(0.010)	(0.0748)
UnemplRate_i1		0.056	0.006	-0.0530
		(0.051)	(0.006)	(0.0399)
lnEPop_ij1		0.489***	0.018***	0.0133
		(0.019)	(0.003)	(0.0174)
distance		-0.694***	-0.045***	-0.0224
		(0.038)	(0.005)	(0.0244)
neighbour			0.010	0.0259
			(0.016)	(0.0526)
colony			0.057*	0.0337
			(0.031)	(0.0737)
lnFreedomPRi1			-0.027**	-0.253***
			(0.011)	(0.0956)
lnFreedomCRi1			0.005	-0.117
			(0.013)	(0.0890)
trend		-0.027***	0.003***	-0.00517
		(0.004)	(0.001)	(0.00346)
Constant	-6.056***	-24.551***	-0.919***	
	(0.059)	(1.505)	(0.226)	
Observations	45458	23374	16221	16221
Adjusted R-squared	0.129	0.597	0.963	0.907

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 5: Diversity in stocks and flows, Estimation of migration flows from 130 countries of origin (i) to 25 OECD destination countries (j), 1985-2006. OLS, FE (destinations and origins) and NLS

	OLS	FE	NLS	NLS	OLS	FE	NLS	NLS	OLS	FE	NLS	NLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(8)	(10)	(11)	(12)
VARIABLES	EFlows	EFlows	EMjit	EMjit	EFlows	EFlows	EMjit	EMjit	EFlows	EFlows	EMjit	EMjit
					-	-	-	-	-	-	-	-
lnHI	0.019**	-0.037**	-0.147**	-0.152	-	-	-	-	-	-	-	-
	(0.009)	(0.015)	(0.0716)	(0.152)	-	-	-	-	-	-	-	-
lnHILang4	-	-	-	-	0.014*	-0.022	-0.124**	-0.249	-	-	-	-
	-	-	-	-	(0.009)	(0.020)	(0.0631)	(0.187)	-	-	-	-
lnsharestockLang4	-	-	-	-	-	-	-	-	0.003	0.040***	-0.00471	0.144***
	-	-	-	-	-	-	-	-	(0.002)	(0.005)	(0.0211)	(0.0340)
EFlows_1	0.857***	0.786***	0.812***	0.724***	0.858***	0.786***	0.824***	0.729***	0.942***	0.878***	0.896***	0.712***
	(0.009)	(0.013)	(0.0340)	(0.0606)	(0.009)	(0.013)	(0.0312)	(0.0548)	(0.003)	(0.007)	(0.0360)	(0.0521)
EStock_1	0.090***	0.132***	0.0918***	0.0249	0.089***	0.132***	0.0706***	0.0230	-	-	-	-
	(0.007)	(0.010)	(0.0247)	(0.0789)	(0.007)	(0.010)	(0.0224)	(0.0771)	-	-	-	-
Destination FE	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Origin FE	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Constant	-8.191***	30.565***	-	-	-8.060***	28.998***	-	-	-4.137***	13.752***	-	-
	(1.386)	(4.810)	-	-	(1.391)	(4.784)	-	-	(1.179)	(4.002)	-	-
Observations	15738	15738	15738	15738	15738	15738	15738	15738	17869	17869	17869	15594
Adjusted R-squared	0.963	0.965	0.908	0.917	0.963	0.965	0.907	0.917	0.960	0.962	0.903	0.918

All tables control for linguistic distance, economic, distance variables and time trend. Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Table 6: Linguistic Diversity in destinations and origins, Estimation of migration flows from 130 countries of origin(i) to 25 OECD destination countries(j), 1985-2006. OLS and NLS

	(1)	(2)	(3)	(4)
Linguistic diversity	OLS EFlows	NLS EMjit	OLS EFlows	NLS EMjit
Measured by: In:	lnElfj4 – a diversity index without distances		lnGIj - a diversity index with distances	
Destination	-0.022*** (0.005)	-0.137*** (0.0449)	-0.014** (0.006)	-0.116** (0.0526)
Origin	0.014*** (0.003)	-0.0136 (0.0128)	0.009** (0.004)	-0.0129 (0.0129)
Observations	*16221*	*16221*	*14815*	*14815*
Adj. R2	*0.963*	*0.910*	*0.964*	*0.909*
	lnPolj4 - a polarization index without distances		lnERj - a polarization index with distances	
Destination	-0.021*** (0.005)	-0.087***(0.013)	-0.087***(0.013)	-0.372***(0.133)
Origin	0.014***(0.004)	0.028***(0.010)	0.028***(0.010)	-0.0765*(0.0445)
Observations	*16221*	*16221*	*16221*	*16221*
Adj. R2	*0.963*	*0.963*	*0.963*	*0.910*
Linguistic diversity in:	lnNoIndLangj – number of indigenous languages		lnNoLangT2P5j - number of languages	
Destination	-0.013*(0.007)	-0.0265 (0.0302)	-0.087***(0.013)	-0.372***(0.133)
Origin	0.012***(0.004)	-0.0944***(0.0189)	0.028***(0.010)	-0.0765*(0.0445)
Observations	*16221*	*16221*	*16221*	*16221*
Adj. R2	*0.963*	*0.911*	*0.963*	*0.910*

The table shows results of regressions with full specification, i.e. we control for linguistic distance, 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    .2215437    .4168986          0        1.5
      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 frika, 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

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 distance:** the index for linguistic closeness between a pair of countries. To be more described

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

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.