The role of language in shaping international migration[*]

(A short form of the title: "Language and Migration")

By Alícia Adserà and Mariola Pytliková

This paper examines the importance of language in international migration from multiple angles by studying the role of linguistic proximity, widely spoken languages, linguistic enclaves and language-based immigration policy requirements. To this aim we collect a unique dataset on immigration flows and stocks in 30 OECD destinations from all world countries over the period 1980–2010, and construct a set of linguistic proximity measures. Migration rates increase with linguistic proximity and with English at destination. Softer linguistic requirements for naturalization and larger linguistic communities at destination encourage more migrants to move. Linguistic proximity matters less when local linguistic network are larger.

JEL Classification: F22, J61, O15

Previous literature has shown that both fluency in the language of the destination country and the ability to learn it quickly play a key role in the transfer of existing human capital to foreign

[*] *Pytliková (corresponding author): VSB - Technical University of Ostrava (VSB-TU), Faculty of Economics, Sokolská tr. 33, 70121 Ostrava (email: mariola.pytlikova@vsb.cz). We are grateful to our anonymous referees and the Editor in charge, for useful comments and suggestions, to participants at NORFACE and CReAM conference on Migration: Economic Change, Social Challenge in London, EEA conference in Oslo, WISE Symposium on Contemporary Labor Economics in Xiamen, IZA AM2 2010 in Bonn, First Tempo Conference on International Migration in Dublin and NORFACE, WB and CReAM conference on "Migration, Development, and Global Issues" in London, and to Bo Honoré for helpful discussions. Jan Bryla, Anne Lund, Enric Boix and Rasmus Steffensen provided excellent research assistance within the migration and linguistic data collection process. We also thank Ignacio Ortuño and Roman Wacziarg for providing the data on linguistic diversity and genetic distance, and Tim Hatton for extrapolated data on migration stocks to the U.S. for years 1980-1989. This research was funded in part by the NORFACE migration program on Migration in Europe - Social, Economic, Cultural and Policy Dynamics (MI3-Migration: Integration, Impact and Interaction), by the Danish Research Council Project No.09-075719 (No.0602-01553B), by the Operational Programme Education for Competitiveness (No. CZ.1.07/2.3.00/20.0296), by the Czech Science Foundation grant (No. GA15-23177S) and by the SGS Research grant (No. SP2014/136). The usual disclaimer applies.*

countries and generally boost immigrant's success in destination countries' labor markets, see Kossoudji (1988), Bleakley and Chin (2004); Chiswick and Miller (2002, 2007, 2010), Dustmann (1994), Dustmann and van Soest (2001 and 2002), and Dustmann and Fabbri (2003). By exploiting differences on adult English proficiency between immigrants from non-English speaking source countries who arrive as young children versus the others, 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. Recent studies show that it is easier for a foreigner to acquire a language if her native language is linguistically closer to the language to be learned (Chiswick and Miller, 2005; Isphording and Otten, 2011). This suggests that the ability to learn and speak a foreign language quickly might be an important factor in the potential migrants' decision.

Besides, a "widely-spoken" native language in the destination country can be a pull-factor in international migration. Two different forces may explain that migration pattern. First, as some "widely spoken" languages are often taught as second languages in schools in many source countries, immigrants are more likely to move to destinations where those languages are spoken in order to lower the costs associated with skill transferability. Second, foreign language proficiency may be valued in the labor market of the source country (European Commission 2002). A recent article by Toomet (2011) finds that knowledge of English is associated with a 15% wage premium in the Estonian labor market. Thus, learning and practising "widely spoken" languages in destination countries may serve as a pull factor especially for temporary migrants.

Although the role of language in international migration is clearly important, this is, to our knowledge, the first paper that disentangles this relationship from multiple angles by studying the role of linguistic proximity, widely spoken language, linguistic communities and language-based policy requirements at destination. Previous evidence on the determinants of migration was limited to including a control for sharing a common language. Only two studies employ some more sophisticated linguistic measures. Belot and Hatton (2012) use the number of nodes on the linguistic tree between two languages to construct a linguistic proximity measure. Likewise, Belot and Ederveen (2012) employ the linguistic proximity index proposed by Dyen et al. (1992) to show that cultural barriers explain patterns of migration flows better than traditional economic variables in a sample limited to developed countries.

In addition we importantly contribute to the literature on determinants of migration by collecting a unique dataset on annual migration stocks and flows for 30 OECD destinations from all world source countries for the period 1980-2010. Moreover, we construct a new set of refined indicators of the linguistic proximity between two languages based on information from the encyclopaedia of languages Ethnologue and relate them to country-pairs on the basis of either the first official, any other official or the major local language in each country.

In the paper, we first use the linguistic indices to examine the relevance of linguistic proximity between origin and destination countries in the decision to migrate and find that emigration rates are higher among countries whose languages are more similar. Migration flows to a country with the same first official language as opposed to one with the most distant language are around 20% higher in models that include a large set of socio-economic and genetic distance controls as well as time and country dummies. The implied differences range from 19 to 35% when using instead either the distance between the major languages in each country or the maximum proximity between any of the official languages and main (if multiple) in both countries. This result is highly robust to the use of two alternative continuous measures of proximity developed by linguists: the *Levenshtein* distance, which relies on phonetic dissimilarity of words in two languages for all world languages, and the *Dyen* index, based on the similarity between samples of words among Indo-European languages. Using these indices the implied increase in emigration rates to countries with similar language as opposite to linguistically distant countries ranges between 14 to 20%. In the context of other determinants of migration, our results show that the effect of linguistic proximity is larger than that of countries that are neighbors or share historical past and of higher (or lower) unemployment rates in origin (or destination), but smaller than the pull effect of income and ethnic networks in destinations. Finally we use the information embedded in our linguistic proximity measure to study whether sharing a particular level of the linguistic tree matters incrementally more than another.

Second, to investigate whether potential migrants prefer a destination with a "widely spoken" language, such as English, as its local language, we estimate separate coefficients on linguistic distance for English and non-English speaking destinations. We find that linguistic proximity matters more for the latter. Migrants' pre-migration exposure to English may temper the relevance of linguistic proximity when studying migration to English-speaking destinations. Further returns to English proficiency may be high in linguistically distant countries and in turn fuel temporary migration from those countries to English-speaking destinations. We explore these models also separately for countries with low educational attainment.

Finally, we find that stricter linguistic requirements for naturalization deter migration flows whereas larger communities with similar linguistic background at destination encourage more migrants to move. These controls do not affect our main results on the linguistic proximity though linguistic distance matters less when the size of the local linguistic network is large.

1. A Model of International Migration

To introduce our empirical specification we present a model of migration across different destinations. This model follows the "human capital investment" theoretical framework (Sjastaad, 1962) and its recent applications in Grogger and Hanson (2011) and Ortega and Peri

(2009). We assume that a potential immigrant maximizing her utility chooses to locate in the country where her utility is the highest among all available destinations.

The utility that migrant *k*, currently living in *i*, attains by moving to *j* is logarithmic and given by:

$$U_{kij} = (y_{kj} - c_{kij})^{\lambda} \exp(\varepsilon_{kij}) \tag{1}$$

where $y_{kj} - c_{kij}$ is the difference between income in destination *j*, y_{kj} (which can be defined in line with Harris and Todaro (1970) as wage times the probability of finding a job, $y = we$), and the cost of migrating from the home country *i* to *j*, c_{kij}.

We can write the probability of individual *k* from country *i* choosing a country *j* among *J* possible destinations as:

$$\Pr(j_k / i_k) = \Pr\left[U_{ijk} = \max(U_{ki1}, U_{ki2}, ..., U_{kiJ})\right] \tag{2}$$

Assuming that ε_{kij} follows an *i.i.d.* extreme value distribution and λ>0, and using the approximation that, $\ln(y_j - c_{ij}) \approx \ln y_j - (c_{ij} / y_j)$, we apply the results in McFadden (1974) to write the log odds of migrating to destination country *j* versus staying in the source country *i* as follows:

$$\ln \frac{M_{ij}}{P_i} = \ln m_{ij} \approx \lambda[\ln y_j - \ln y_i] - \lambda C_{ij} = \lambda[\ln w_{kj} + \ln e_{kj} - \ln w_{ki} - \ln e_{ki}] - \lambda C_{ij} \tag{3}$$

where M_{ij} are flows of individuals from *i* to *j*; P_i are the stayers; m_{ij} is the emigration rate from *i* to *j* and C_{ij} are migration costs expressed as a proportion of destination income, $C_{ij} = (c_{ij} / y_j)$

.

The probability of migration depends on the difference between income related to staying at home country i or migrating abroad j adjusted for costs of migration (e.g. psychological and direct out-of-pocket costs and those associated with imperfect skill transferability).

Equation (3) relies on the assumption that the relative probabilities of two alternative locations only depend on the characteristics of those two alternatives. Since the empirical analysis of our paper includes only OECD destinations, we only need that the IIA holds for these countries (McFadden, 1974; Grogger and Hanson, 2011). As a way to test the IIA assumption across OECD destinations, we re-estimate models by excluding one destination at a time. Results are stable and thus suggest that such an assumption is plausible here.[1]

2. Data Construction

2.1. International migration data

For the analysis we have constructed a new dataset on immigration flows and stocks of foreigners in 30 OECD destination countries from 223 source countries for the years 1980–2010. The dataset was collected by writing to selected national statistical offices for the majority of the OECD countries to request detailed yearly information on immigration flows and foreign population stocks by source country in their respective country. For three countries, Korea, Mexico and Turkey (and partly Japan), we obtained the data from the OECD International Migration Database. The first version of the migration dataset covered 22 OECD destination and 129 source countries over the period of years 1989-2000, see Pedersen, Pytlikova and Smith (2008). For the purpose of this paper we added data from eight additional destination countries – Czech and Slovak Republics, Hungary, Poland, Ireland, Turkey, South Korea and Mexico -

[1] Results are available from authors.

and extended the number of countries of origin to cover the entire world. Further, we prolonged the time period covered by the data to include the years 1980-1989 and 2001-2010.

Our international migration data set presents substantial progress over that used in past research on determinants of migration.[2] First, our data covers annually both migration flows and foreign population stocks.[3] Second, the data is more comprehensive with respect to destinations, origins and time due to our own effort with data gathering from particular statistical offices. For an overview of comprehensiveness of observations of flows and stocks across all destination countries over time, see the Appendix Table A1 and Table A2, respectively. It is apparent that the data becomes more comprehensive over time and thus missing observations become less of a problem for more recent years. In our dataset, as in the other existing datasets, different countries use different definitions of an "immigrant" and draw their migration statistics from different sources[4]. In particular for foreign population stock, we preferably use the definition

[2] See data by Docquier and Marfouk (2006), OECD (2011), the World Bank (2011), and the United Nations (2011).

[3] Migration flow is the inflow of immigrants to a destination from a given origin in a given year. The definition usually covers immigrants coming for a period of half year or longer. Foreign population stock is a number of foreigners from a given country of origin living in a destination in a given year. The foreign population stock data is dated ultimo.

[4]Thus our data set bears some problems related to different sources of migration data (censuses, registers or labor force surveys), different definitions of foreigner (country of birth and citizenship) and unbalanced nature of the data due to missing observations for some countries of destinations and origins. For example, Austria, 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 issued; Australia, Canada, New Zealand and Poland use data from censuses; some countries like Greece, the United Kingdom and the United States use labor force surveys and others have information based on social security systems or other sources. 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

based on country of birth to determine the origin of migrants. See the Appendix Tables A3 and A4 for a detailed overview of definitions and sources for data on immigration flows and foreign population stock, respectively.

2.2. Linguistic distance

We use three different linguistic distance indices for our analyses: (1) a newly constructed *Linguistic Proximity* index, described below and based on information from Ethnologue, (2) the *Levenshtein distance* developed by the Max Planck Institute for Evolutionary Anthropology and (3) the *Dyen linguistic proximity* measure proposed by Dyen et al. (1992).

We create a measure that captures the linguistic proximity between two languages based on information from the encyclopaedia of languages Ethnologue (Lewis, 2009). The *Linguistic Proximity index* ranges from 0 to 1 depending on how many levels of the linguistic family tree the languages of both the destination and the source country share. To construct the index we first define a set of increasing weights: the first equal to 0.1 if two languages are related at the most aggregated linguistic tree level, e.g. Indo-European versus Uralic (Finnish, Estonian, Hungarian); the second equal to 0.15 if two languages belong to the same second- linguistic tree level, e.g. Germanic versus Slavic languages; the third equal to 0.20 if two languages belong to the same third linguistic tree level, e.g. Germanic West vs. Germanic North languages; and the fourth equal to 0.25 if both languages belong to the same fourth level of linguistic tree family, e.g. Scandinavian West (Icelandic) vs. Scandinavian East (Danish,

population differs among countries as well, but for the majority of destinations we use the definition by country of birth. Australia, Austria, Canada, Denmark, Finland, France, Iceland, Ireland, Mexico, New Zealand, Norway, Poland, the Slovak Republic, Spain, Sweden, Turkey, the United Kingdom and the United States define immigrant stock by country of birth. A few countries like Belgium, Czech Republic, Germany, Greece, Hungary, Italy, Japan, Korea, Luxembourg, the Netherlands, Portugal and Switzerland define immigrant population by citizenship.

Norwegian and Swedish), German vs. English, or ItaloWest (Italian, French, Spanish, Catalan and Portuguese) vs. RomanceEast (Romanian). Then, we construct the linguistic proximity index as a sum of those four weights to capture the maximum number of shared linguistic family tree's branches, and we set the index equal to 0 if two languages do not belong to any common language family, and equal to 1 if the two countries have a common language. Thus the linguistic proximity index equals 0.1 if two languages are only related at the most aggregated level of the linguistic, 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 up to the third linguistic tree level, e.g. Germanic North languages; and 0.7 if both languages share the first four levels, e.g. Scandinavian East (Danish, Norwegian and Swedish). We use this measure to build a matrix that contains metrics of proximity between any pair of languages from our destination-source pairs and provides a better adjusted and smoother indicator of proximity than the standard dummy for common language used in most of the literature. To link the linguistic proximity measure to country pairs we initially use the first official language in each country.

Figure 1 shows the distribution of the linguistic proximity among country pairs employed in the baseline analysis of the paper and for which we do not have missing observations in the control variables – the distribution is essentially the same without this restriction. Around 42.5 per cent of country-pair observations do not share any branch of the tree and 36 per cent are only related at the most aggregated level. Only in around 4 per cent of the observations do both countries share the same language, whereas the proportions of observations that are related at the second, third and fourth level of the linguistic tree stand at 8 per cent, 6.7 per cent and 2.7 per cent, respectively.

Figure 2 presents more detailed information about the distribution of migration flows by the linguistic tree level the origin and destination country share. As seen in Table 1, on average, a total number of about 1,077 people migrate from specific origin to another OECD destination country per year. During the period of 1980-2010, there were in total about 110 million people migrating to another OECD country: among them about 14.6 million people migrated to countries that share the same first official language and about 40 million migrated to countries whose first official languages did not have any level in common with that of their country of origin. The largest proportion of migrants, around 45 million, migrated to countries whose languages share only the most aggregate linguistic tree family, and about 1.6, 7 and 2.1 million to countries sharing the second, third and fourth level of the linguistic tree, respectively. The overall pattern is not that different when looking at flows by major language spoken, though more migrants are moving to countries with major languages very distant from theirs. When all official and main languages are considered, the flows to destinations with a common language are strikingly large. Of course this is in part due common colonial past, which we take into account in our empirical specification.[5]

In addition to our index, we employ two continuous measures of linguistic distance between countries developed by linguists. The first one is the Levenshtein linguistic distance produced by the Max Planck Institute for Evolutionary Anthropology, which relies on phonetic dissimilarity of words in two languages. Linguists choose a core set of the 40 more common words across languages describing everyday life and items; then, express them in a phonetic transcription called ASJP code and finally compute the number of steps needed to move from one word expressed in one language to that same word expressed in the other language. For a

[5] When we split the sample by decades (not shown here), we observe that flows are increasing over time and that, despite our panel is somewhat unbalanced, this change is pretty proportional across all linguistic tree levels.

detailed description of the method, see Bakker et al. (2009).[6] In our country sample the index ranges from 0 (when the two languages are the same) to a maximum of 106.39 (for the distance between Laos and Korea). The second one is 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. The index increases with similarity between languages and it is equal to 1000 when the two languages are the same.[7]

The correlation between all linguistic indices is above 0.9, while the correlation of these indices with measures of the genetic distance of the population of two countries, a basic control included in our models below, ranges only between 0.13 and 0.06.

3. Empirical Model Specification

On the basis of equation (3), our econometric model assumes that emigration rates to one destination are driven by differences in wages, employment rates between origin and destination countries, and the costs of migration:

$$\begin{aligned}\ln(\boldsymbol{m}_{ijt}) &= \gamma_1 + \gamma_2 \ln(\boldsymbol{gdp}_{jt-1}) + \gamma_3 \ln(\boldsymbol{gdp}_{it-1}) + \gamma_4 \ln(\boldsymbol{u}_{jt-1}) + \gamma_5 \ln(\boldsymbol{u}_{it-1}) + \gamma_6 \ln(\boldsymbol{pse}_{jt-1}) + \\ &+ \gamma_7 \ln(\boldsymbol{s}_{ijt-1}) + \gamma_8 \boldsymbol{L}_{ij} + \gamma_9 \boldsymbol{D}_{ij} + \gamma_{10} \boldsymbol{FH}_{it-1} + \gamma_{11} \boldsymbol{lr}_{jt-1} + \gamma_{12} \ln(\boldsymbol{p}_{ijt-1}) + \boldsymbol{\delta}_j + \boldsymbol{\delta}_i + \boldsymbol{\theta}_t + \boldsymbol{\varepsilon}_{ijt}\end{aligned} \tag{4}$$

where $\boldsymbol{m}_{ijt}$ denotes gross flows of migrants from country i to country j divided by the population of the country of origin i at time t, where $i=1,\dots,223;\ j=1,\dots,30$ and $t=1,\dots,31$. As in previous studies we proxy wages by GDP per capita and employment prospects in the

[6] The Levenshtein linguistic distance has been used, for example, to measure the difficulty in learning the local language among migrants to Germany (Isphording and Otten 2011, 2013). To the best of our knowledge we are the first to employ the Max Planck index in analyses of migration determinants.

[7] For application of the Dyen index in the context of determinants of migration, see Belot and Ederveen (2012).

sending and receiving countries by unemployment rates, $\boldsymbol{u}_{jt}$ and $\boldsymbol{u}_{it}$. In some models we introduce the level of GDP per capita in the source country in a quadratic form, $\ln(\boldsymbol{gdp}_{it-1})^2$, as a means to test for the non-linearity effects found by previous works (Chiquiar and Hanson 2005, Hatton and Williamson, 2005 and 2011, Clark et al. 2007, Pedersen et al. 2008; Docquier and Rappaport, 2012; Belot and Hatton, 2012). The hypothesis behind this line of work is that extreme poverty constrains the ability to cover costs of migration, and as income levels rise beyond extreme poverty, migration increases. However after GDP reaches a certain level, migration could again decrease because the economic incentives to migrate to other countries decline.[8] In addition, Borjas (1999) argues that generous social security payment structures may play a role in migrants' decision making. Potential emigrants take into account both the probability of being unemployed and the generosity of welfare benefits in the destination country that constitute a substitute of earnings during the period of job search.[9] We include public social expenditure as percentage of GDP, $\boldsymbol{pse}_{jt-1}$, as a proxy for the "welfare magnet" among explanatory variables.

We expect costs associated with migration to be larger with physical, cultural and linguistic distance between countries, but to fall with the existence of migration networks (i.e. networks of family members, friends and people of the same origin that already live in a host country). In addition migration costs may depend on specific destination and origin factors (such as immigration laws in destinations or credit-market constraints at origin). In our empirical specification we use the total foreign population from country *i* living in country *j* per

[8] Lack of good South-South migration data may also account for this finding if individuals from the poorest countries migrate to close and relative poor countries.

[9] In fact a similar argument would warrant the inclusion of social expenditures in the country of origin into our empirical model. Unfortunately, as in previous research, data constraints preclude us from including this information as only the OECD provides good comparative data on social expenditures.

population of the source country *i*, s_{ijt}, to control for the network of migrants that has been shown to play an important role in lowering the direct and psychological migration costs (Massey et al., 1993; Munshi, 2003). Additionally in the robustness analyses we control for the total stock of migrants with the same linguistic background as a migrant from a particular origin to be able to account for the effect of linguistic enclaves on the propensity to migrate.

Matrix $\boldsymbol{L}_{ij}$ includes measures of linguistic distance between countries described in IIIB to test the main hypothesis of the paper, namely whether larger language barriers increase migration costs for an individual by setting hurdles to the transfer of her skills and her integration in the receiving society.

Even with recent improvements in communication technologies, the continued globalization of the economy and declining costs of transportation, physical and cultural distance are bound to raise the direct cost of migration. To control for the effect of distance, matrix $\boldsymbol{D}_{ij}$ includes the following variables: *Log Distance in Kilometres* between the capital areas in the sending and receiving countries; *Neighbor Country* which takes a value of 1 if the two countries are neighbors; and *Historical Past Dummy,* with value 1 for countries ever sharing historical past. Past common history might decrease the cultural distance between countries, and increase the information available about the potential destination country. Further, we include measures of the *genetic distance* between populations of both countries in our regressions as an additional control for cultural factors that could be confounded with our linguistic distance indices,. These indices, provided by Roman Wacziarg, are based on the work by Cavalli-Sforza, Menozzi, and Piazza (1994) and have already been employed in other contexts to study, for example, cross-country differences in development (Spolaore and Wacziarg 2009). A detailed explanation of how the indices were constructed can be found in these two publications. The ”dominant” genetic distance measures, for each pair of countries, the distance between the ethnic groups with the largest shares of population in each country. It increases with the differences between two populations and takes a zero if the distributions of alleles in both populations are identical. In one model specification, we also include *Log Trade Volume*, which is defined as the (log) total trade values (both imports and exports) for all country pairs. The import and export value are collected from the Direction of Trade Statistics and are expressed in nominal US dollar prices. We expect that the business ties represented by the volume of trade to have (positive) effects on international migration. Moreover, this variable is often considered as an indicator of globalization and cultural proximity.

Matrix $\boldsymbol{FH}_i$ includes a couple of indices from *Freedom House,* which aim to separately measure the degree of freedom in political rights and civil liberties in each country. Each variable takes on values from one to seven, from the highest degree of freedom to the lowest. Violated political rights and civil liberties may increase migration outflows in a given country. On the other hand, political restrictions may also impede outmigration.

Further to account for differences on how much language matters for policy in each country, we build a time varying index on language requirements for naturalization in each destination $\boldsymbol{lr}_{jt-1}$ as detailed in the results section. Finally we include a variable that captures the relative population size in destination with respect to origin, $\boldsymbol{p}_{ijt-1}$, in order to control for demographic developments.

All variables used in the estimations, except dummy variables and the linguistic proximity indices, are expressed in logarithms. Table 1 contains definitions, sources and summary statistics of all variables. In order to account for what information was available to the potential migrant at the time the migration decision was made, the relative differences in economic development and employment between origin and destination countries are lagged by one period. More importantly, there might be a problem of reverse causality if migration flows impact both earnings and employment. Lagging the economic explanatory variables and treating them as predetermined is one way to reduce the risks of reverse causality in the model.

Since the stock is just a function of previous stock plus migration flows minus out-migration, we also lag it and assume that the lagged stock is predetermined with respect to the current migration flows.

All specifications contain a set of year dummies, θ_t, in order to control for common idiosyncratic shocks over the time period and robust Hubert/White/sandwich standard errors clustered at each pair of destination and source countries. Models also contain country of destination and country of origin fixed effects, δ_j and δ_i, separately to capture unobserved characteristics of immigration policy practices in each destination country, credit market constraints in origins, as well as climate, openness towards foreigners or culture in each country, among other things. Some previous literature includes pair-wise fixed effects to capture (unobserved) traditions, historical and cultural ties between a particular pair of destination and origin countries. We cannot include them here since they would be collinear with our linguistic proximity variables of interest. However, as described above, we use a number of explanatory variables that help to control for the historical and cultural ties between countries.

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 (only around 4.5 % in our data).[10] Even though most previous studies on migration determinants have used linear models with log-transformed variable, a few have chosen count models to fit the nonnegative dependent variable (e.g. Belot and Ederveen (2012) use negative binomial; Simpson and Sparber (2013) use Tobit and Poisson count models). In Table 2 below we present Poisson estimates of our baseline model in column (9).[11]

[10] This percentage is much lower than either the 95% of zero values that Simpson and Sparber (2013), who specifically discuss the "zero problem" in migration data, face or the usually reported in the trade literature when estimating gravity models.

[11] We obtained similar estimates of the model using nonlinear least squares where the level of migration flows is explained by the exponential of the linear combination of all log-transformed independent

4. RESULTS

4.1. Linguistic proximity

Table 2 first shows results from the most parsimonious model that only includes the linguistic proximity index and a constant to the full specification. Columns (1) and (2) in Table 2 show that our linguistic proximity index alone accounts for 11% of the variance in world migration rates whereas the common language dummy used in the previous literature only explains around 7%. When both are included in the same model in column (3), only the linguistic proximity index is significant in a sample that encompasses around 100,000 observations.

The remaining columns in Table 2 include the basic pull and push factors as well as country and time fixed effects as given by equation (4). The coefficient of linguistic proximity is positive and highly significant in all specifications. Thus, other things being equal, emigration flows between two countries are larger the closer their languages are. As expected, the coefficient decreases in size as more controls are added to the model; in particular it shrinks from around 0.73 to around 0.2 when migrant's stocks are included in columns (7) to (12). The latter suggests that a large network from the same origin may alleviate the pressure of learning the local language to assimilate to the new labor markets and society. We test the effects of "linguistic networks" on migration and linguistic proximity in depth in subsection D. Column (6) adds measures of social expenditure in destination as well as unemployment rates in origin and destination. Source country unemployment rates impose the largest restriction with respect to the number of missing observations. By including unemployment variables the number of observations halves. In order to take advantage of the large sample of migration flows and stocks we have gathered, for the remaining models we create an alternative measure of

variables without imposing any restrictions between the mean and the variance as some count models require.

unemployment where we substitute missing observations by the mean unemployment over the sample and include dummy indicators where unemployment is missing (columns 7 to 12).

The coefficient of linguistic proximity in the model with migration stocks and unemployment in column (8), our baseline specification for the rest of the paper, is 0.209 and it is highly significant at 1%. It implies that emigration flows to a country with the same language as opposed to a country with the most distant language should be around 20% higher, ceteris paribus. Thus emigration rates to France from Benin where French is the first official language should be around 18% higher than those from Zambia's with a linguistic index of 0.1 with respect to France or 6% higher than those from Sao Tome, with a linguistic index of 0.7, ceteris paribus.[12]

Similarly to previous studies such as Bauer et al. (2005), Clark et al. (2007), Pedersen et al. (2008), McKenzie and Rapoport (2010) and Beine et al. (2011) we find network effects to be an important determinant of subsequent migration. Results in column (8) indicate that a 10% increase in the stock of migrants from a certain country is associated with an increase of around 6.7% in the emigration rate from this country, ceteris paribus.[13] Further a 10% increase in the GDP of the destination country is associated with an increase in emigration rates to that country of around 17%. While the level of GDP per capita in origin enters negatively and significantly in column (4) as expected, we test for the presence of nonlinearities found by the previous literature in the remaining columns of Table 2. Estimates in columns (5) and (7) conform to the hypothesis that emigration rates increases with per capita income for very low levels, in this case up to annual levels of income below $800, and then decrease as the economic incentives

[12] As an additional test on whether the IIA assumption holds for OECD destinations, we restrict the data to study migration flows among OECD countries. The coefficient of the linguistic proximity index remains highly significant at 1% and increases to around 0.39.´Results are available from authors.

[13] If t-2 lags in the migration stock variable are used instead, its coefficients are slightly lower, but the rest remains unchanged. Results are available from authors.

to migrate diminish.[14] Once the stock of migrants from the country of origin is included in the models, migration rates clearly decrease with income per capita in the source country. A possible explanation for this finding is that the networks of friends and ethnic fellows can help to alleviate the poverty constraints to migrate as suggested by, for example, Hatton and Williamson (2002, 2011), and Pedersen et al. (2008).

Emigration rates are significantly higher from countries with relatively high unemployment rates and lower to destinations with high unemployment, other things being the same. In line with the theoretical framework proposed by Borjas (1999), we find that the coefficients to public social expenditure are positive and significant. This runs contrary to some existent empirical evidence (Zavodny 1997, Pedersen et al. 2008 and Wadensjö 2007, among others) and more in line with other works reviewed by Guiletti and Wahba (2013). At any rate social expenditures would only be relevant for migrants as long as they are entitled to receive them, but some of the OECD countries provide universal benefits to anybody eligible regardless of nationality.[15] Population ratio, common history and shorter distance are significantly associated with stronger emigration flows. In our preferred specification in column (8), having a past historical tie increases the emigration rates to a destination by around 26%. Lack of political liberties seems to increase outmigration, but coefficients fail to be significant in most specifications. Conversely, controlling for political rights, emigration rates are larger from countries with better civil rights. Some of these rights are associated with lower barriers to out-migration and geographic mobility. Finally findings regarding linguistic proximity are robust to the inclusion of measures of genetic distance which are only significant and negative (as expected) in models that do not include stocks (columns 4 to 7), but not once network controls

[14] In column (6) that encompasses the smaller sample for which unemployment rates are available in origin, emigration rates increase, though very moderately, with GDP at origin.

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

are added. 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.[16] Our findings are also robust to the inclusion of bilateral trade volume, in column (12), which is often considered an indicator of globalization and cultural proximity. The volume of trade is associated with larger migration flows.

To sum up, we find that linguistic proximity has an important role in migration. Sharing the same language versus not sharing any level of the linguistic family tree has an effect on immigration flows equivalent to an increase of 12% in destination country GDP. To get a better sense of the relative importance of the linguistic proximity compared to other pull and push factors, we include in column (9) the standardized beta-coefficients of column (8) which give us a measure of the changes in standard deviations of migration rates resulting from a change in one standard deviation of each factor. This is of course not a perfect measure since we may attach very different meaning to similarly relatively sized changes across different factors. An increase in one standard deviation in the existing stock of migrants is associated with a 0.76 standard deviation increase in migration rates. A similar increase in the income per capita of the destination country increases migration to this country by 0.2 standard deviations, whereas the implied impact of linguistic proximity is just a tenth of that, around 0.02 standard deviations. At any rate the impact of having closer languages is larger than that of countries having higher (or lower) unemployment rates in origin (or destination) but less than half of the pull implied from larger social expenditures in destination.

[16] Results are also robust to the use of a second index ("weighed") that 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. Estimates with this alternative "weighed" genetic distance are available from the authors upon request.

4.2. Robustness

To test the robustness of our results, in Table 3 we use a set of alternative measures of linguistic distance. The first three columns in Table 3 present the baseline model estimated first with our index of linguistic proximity, and then, in column (2) with the Levenshtein index (divided by 100) and in column (3) with the Dyen index (divided by 1000), all calculated for the first official language in each country. Given that the Levenshtein index is defined in terms of distance as opposed to proximity between languages, the significant negative estimate in column (2) indicates that emigration rates are larger to countries with closer languages. The coefficient implies that emigration rates to countries with similar languages should be around 15% higher than to those with an index of around 100 (quite dissimilar). Similarly, in column (3) the coefficient for the Dyen index is significantly positive and implies that emigration rates from a country with the same language (and a Dyen index of 1000) are around 18% larger than those from a country with a rather dissimilar language (the minimum of around 100 in our sample of Indo-European languages). Because the Dyen index covers only Indo-European languages, our number of observations in the most complete models presented in the paper is reduced significantly from over 51,000 to only close to 28,000. It is interesting that the size of the implied effect with the Dyen is remarkably similar to those obtained with the other two indices, even though the sample is smaller and restricted to countries that are likely more homogenous. The coefficient of 0.203 in column (3) implies that 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 17%. 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 should be in order of 15% and 8.5% respectively.[17] Table 3

[17] In separate estimates, we use the Dyen index and attach a zero value for the pairs of countries in which only one of the languages is Indo-European. The estimated coefficient on this linguistic index of 0.146

includes one row with standardized coefficients for each model. They confirm the closeness of results across the three linguistic indices since one standard deviation change in each of them results in between 0.022 and 0.013 standard deviation changes in migration rates.

Next, we extend the set of linguistic measures to take into account the existence of multiple official (first and second most-spoken, even if not official) languages and we calculate the maximum proximity between two countries using any of those languages. Figure 1 shows that the distribution for this index when measuring the linguistic proximity with the Ethno-linguistic tree shifts towards more closeness between countries compared to the distribution when only first official languages are taken into account. Using all official and main languages, the percentage of country-pair observations that have no branch in common shrinks to 22, while both countries share at least one common official language in 10 per cent of the cases. The proportions are also higher at all other levels: over 38 per cent at level 1; 14 per cent at level 2; 8 per cent at level 3 and 7 per cent at level 4. The main reason for this change is that many former colonies retain the official language of their colonizing power (i.e. English, French, and Portuguese) among their official 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. Finally, with the same methodology we construct an index of linguistic proximity using instead the language most extensively used in the country (the "major" language) even if in some countries it is not among the official ones. Not surprisingly,

is, not surprisingly, slightly smaller than when the sample is restricted to only Indo-European countries (but still significant). Conversely, when we restrict the sample to countries with Indo-European languages, the estimated coefficient of our index of linguistic proximity is larger than that obtained in Tables 2 and 3.

the linguistic proximity index is equal to zero for the majority of country pairs (58 per cent) and around 28 per cent only share the first branch of the linguistic tree. Just 2 per cent of the observations share a common major language and the proportions for the other levels are also lower than for first official languages (2.25 per cent at level 4, 4 per cent at level 3 and 5 per cent at level 2). The coefficients of the linguistic proximity when using the two alternative criteria, shown in columns (4) to (9) of Table 3, are significant and positive. Yet the size of coefficients reveals some small differences with respect to indices based on the first official language. For our linguistic proximity index, the explained differences in emigration rates when using all official and main languages in column (4) and, particularly when using major languages in column (7) (where the standardized coefficient increases to 0.027), are a bit larger. When using the Levenshtein index, the coefficient is the largest when using major languages in column (8) (in part due to the fact that languages tend to be more dissimilar on average) but the standardized coefficient is the largest when taking into account all official and main languages in column (5). When looking at results with the Dyen it is important to understand that the size and composition of the sample varies tremendously. In fact the sample size when using all official languages in column (6) is twice as large as when using the major language in column (9), since many of the latter are not Indo-European. The standardized coefficient is the largest in column (6), where a one standard deviation change in the Dyen implies an increase of almost 0.04 standard deviations in migration rates.[18]

[18] As an additional robustness analysis we took advantage of the detailed information embedded in our linguistic proximity index and substituted (in the baseline model) the linguistic proximity index for an indicator of whether the languages of two countries share a particular level in the Ethno linguistic tree. We estimated the model separately for each level to see whether there are some non-linearities in the relevance of linguistic proximity. Dummies for all levels of the linguistic family tree - except for the most aggregated (Indo-European vs. Uralic) and the second most aggregated (Germanic vs. Slavic)– display a significant positive coefficient that increases up to the fourth level of the tree. When the

4.3. The Role of Widely Spoken Languages

Our linguistic proximity index does not capture the importance of some widely spoken Indo-European languages (particularly English) in the media, in business or as a choice of second language in schools (see Eurobarometer study on languages by European Commission 2006). Further, if foreign language proficiency is an important part of human capital in the labor market of source countries (see European Commission 2002 on language proficiency as an essential skill for finding a job in home countries), returns to proficiency in widely spoken languages may be particularly high in countries which are linguistically distant from the widely spoken language. Thus learning, practicing, and improving the skills of "widely spoken" languages in "native" countries could serve as a pull factor especially for temporary migrants who take this skill back home. Models in Table 4 include separate indicators of linguistic proximity for non-English and for English speaking destinations to examine the role of English as one of the most widely spoken languages in the world. If there is some advantage from knowing English as a second language, we expect that the linguistic proximity between native languages should matter more for non-English speaking destinations than for the others. Results in Table 4, column 1, show that the linguistic proximity index is a strong predictor of emigration rates toward non-English speaking destinations. The coefficient for English destinations is much smaller than that for non-English destinations, and it is statistically insignificant. This is consistent with the hypothesis that people migrate to destinations with a widely spoken language even if their mother languages are linguistically far from that language. The finding is similar in column (2) when we use the linguistic proximity of the major language in the country instead. As a matter of fact, the coefficient to non-English speaking destination is even larger that when the first official language is employed in column (1). In column (3) we use the

analysis is restricted to the last fifteen years level 2 is also significant. The Table with those additional robustness tests is available from the authors upon request.

proximity index for the closest pair among all official and main languages of each country. The coefficient for English destinations is now slightly larger and statistically significant at the 10 per cent level, but still lower than that for non-English speaking destinations. This is probably related to the fact that English and other colonial languages are (when not first) likely second or third official languages in many countries where they are not necessarily neither majoritarian nor widely known by the whole population but are taught in schools.[19]

Broadly speaking, research based on micro-data unveils two polar types of migrants: on the one hand, low skilled manual workers in jobs that are not filled by the natives in the destination country and, on the other hand, high skilled professionals (in IT or science, among other fields) (see Belot and Hatton 2012; Docquier and Rappaport 2012 for an overview). Since language plays a key role in a successful transfer of immigrants' home country education and skills to foreign labor markets (Kossoudji, 1988; Bleakley and Chin, 2004; Chiswick and Miller, 2002, 2007 and 2010; Dustmann, 1994; Dustmann and van Soest, 2001 and 2002; and Dustmann and Fabbri, 2003), the relevance of linguistic proximity and knowledge of widely spoken language will likely differ across these various groups of migrants with different needs of skill transferability. Ideally we would like to have individual level data in order to explore these (likely) heterogenous effects of language, but unfortunately a dataset at the level of the individual migrant is not currently available. Keeping in mind potential problems of ecological inference driven by differential selection across countries (Docquier et al. 2007), the large number of destination and origin countries in our dataset allows us to analyze the migration patterns for groups of countries with different levels of educational attainment.

In columns (4) to (6) in Table 4 we restrict the sample to countries in the lowest quartile in gross secondary education enrollment rates by year. Results seem to support the hypothesis that

[19] Results are robust to including measures of the number of computers per capita in the source country to infer exposure to English or other languages through internet.

linguistic proximity and knowledge of a widely spoken language are less relevant for migrants with lower average skills. The coefficient for non-English destinations in column (4) is lower than in column (1) and barely significant, whereas the coefficient for English-speaking destinations remains statistically insignificant. When we use linguistic proximity between major languages, we find no effect of linguistic proximity for both English and non-English speaking destinations. Conversely, both coefficients are positive, significant and larger than in column (3) when we employ the closest pair among all official and main languages of each country.[20] With lack of individual data it is difficult to tease out the competing mechanisms behind this finding. A large proportion of countries in this restricted sample are former colonies and some of the widely spoken languages in the world turn up among their official languages. Thus, even if potential migrants do not learn those formally as an additional foreign language, they may have received basic instruction in English (or other major languages). In fact among the country-pairs with an English speaking destination in this restricted sample, close to 60% of the observations share English as a common official language (as opposed to 40% in the complete sample). Further, a positive selection within those countries will make the average migrant more likely to have received some education in English.

In columns (7) and (8), we add tertiary enrollment rate into our model as a proxy for country's level of education and we find that for the sample of all countries, those origins with higher tertiary enrollment rates have larger migration outflows. This is in line with the human capital investment theoretical framework prediction that more educated individuals are more mobile. The relevance of linguistic proximity is robust to the inclusion of tertiary education in column

[20] In fact the effect implied by standardized coefficients is also larger than when using the entire dataset. The corresponding beta coefficients for linguistic proximity based on the closest pair among all official languages in column (3) are 0.022 and 0.015 for non-English and English speaking destinations, respectively, whereas they almost double, to 0.044 and 0.026, when using the sample of countries with low secondary enrolment in column (6).

(7) and it increases with the level of gross tertiary enrollment in column (8). The latter likely exposes an increasing need for skill transferability.

4.4. Linguistic Networks and Policy

A potential concern when interpreting our findings is to what extent linguistic proximity interacts with migration policy, which has been shown to be an important factor to explain migration flows (Mayda 2010, Ortega and Peri 2009). We contacted several specialists in the field and went over legislation in an attempt to gather a comparable index to account for the strictness of language requirements at entry. Given the heterogeneity of schemes across countries (skilled and unskilled workers; economic, spouse or student visas, among others),[21] we followed the advice of those experts who suggested naturalization policy requirements would be easier to measure in a homogeneous way. We have combined existing information gathered by previous research (Goodman 2010a, Weil 2001, Waldrauch 2006, Joppke 2007, country official websites and data from the project EUDO Citizenship Observatory, among others) and we have read all the pertinent legislation on citizenship by country available in the eudo-citizenship.eu website. We have created a time-varying index that measures whether countries have any language requirement in the naturalization process, whether the requirement is formal (i.e. written test) or informal and whether it has changed in each of the 30 OECD destinations for the 1980-2010 period. We include this index in the models in Table 5 (columns 1 & 2). Results show that, ceteris paribus, migration flows are smaller in countries with higher

[21] When we used some basic information on entry requirements gathered by Goodman (2010b) for 15 EU countries, only available for a couple of years of the sample, our estimates on the effect of linguistic proximity were not altered. We also employed some general country classification on whether selective point systems were used or not in some countries (many of them English speaking). Results showed that linguistic distance mattered as much (in fact a bit more) in those countries without a point system.

linguistic requirements for naturalization, but the coefficient of linguistic proximity and its significance remain unaltered, even when we include its interaction with the policy requirement in column (2).

In addition to providing information and affective support for the newcomer, a large stock of migrants from the same origin is likely associated with lower pressure to learn the local language immediately after arrival as it facilitates the existence of "language enclaves" within destination countries. In that regard, the relevant community for a newcomer may be the one composed by individuals that share the same linguistic background. We have constructed two indicators that measure the size of the linguistic networks: the total stock of migrants that share the same level of the linguistic tree (either at level 3 or at level 4). The coefficient of those indicators in columns 3 & 4 of Table 5 imply that a larger linguistic community significantly attracts more migrants to a destination. In addition we include the interaction of the size of the linguistic network with the index of linguistic proximity. The coefficient is negative and significant indicating that linguistic distance matters less when the size of the linguistic community (potentially, the linguistic enclave) is large in the destination country.

5. Conclusions

In this paper we construct a new dataset on migration flows into and stocks of foreigners in 30 OECD destination countries from 223 source countries for the years 1980–2010 in order to study the role of language in shaping international migration. Specifically, we investigate how linguistic distance, the presence of a widely spoken language at destination, linguistic requirements in immigration policy and the existence of linguistic enclaves in destinations are related to migration flows. Besides collecting the largest international migration data set to date, we construct our own linguistic proximity measure, based on information from the encyclopaedia of languages Ethnologue. We find that migration rates are higher between countries whose first official languages are closer. The result holds when we instead use either the proximity between the most commonly used language in each country or the minimum distance between any of the multiple official and main languages in both countries. This finding is also highly robust to the use of two continuous distance measures developed by linguists (the Dyen and the Levenshtein indices) and to the inclusion of a number of variables that capture cultural, historical and trade ties between countries, such as genetic distance, dummies for common historical past and common border, distance, and bilateral trade ties. This suggests that language itself affects migration costs beyond the effects of cultural homogeneity or physical proximity between origin and destination countries. In the context of traditional economic push and pull factors found in the literature, the impact of linguistic proximity on migration flows between two countries is lower than that of ethnic networks or destination GDP per capita level, but much stronger than that of unemployment rates.

To investigate the role of English, a widely spoken language, in migration, we estimate separate coefficients on linguistic proximity for English and non-English speaking destinations and we find that linguistic proximity matters more for the latter group. Pre-migration exposure to English by the average migrant probably weakens the relevance of linguistic proximity indicators to English speaking destinations. We also find that linguistic proximity (particularly to English-speaking destinations) is less relevant for migrants coming from countries with low levels of secondary enrollment. Overall, there is more emigration from countries with higher levels of tertiary education, and the importance of linguistic proximity increases with the level of tertiary education at origin. This may reflect the increasing need for skill transferability for highly skilled migrants.

Finally, we investigate the role of immigration policy and linguistic enclaves on migration. Immigration policy with stricter requirements of language proficiency may affect migration flows and the impact of linguistic proximity. To test this, we create a time-varying language requirement index for naturalization for our 30 destinations for 1980-2010. Results show that even though migration flows are smaller in countries with higher requirements, the relevance of linguistic proximity remains unaltered. Further, migration rates are significantly larger in destinations with larger size of the linguistic community, where the pressure to learn the local language immediately after arrival is likely to be lower. Our estimates reveal that the linguistic proximity matters significantly less when the size of the linguistic community (the linguistic enclave) is large in the destination country.

Our research contributes to the understanding of the determinants of the direction of migration flows across countries and highlights the importance of migration costs as obstacles to greater international migration. Governmental policies aimed at promoting instruction of foreign languages both at origins and destinations can foster the recruitment and mobility of international workers. Some existing empirical literature suggest that the presence of foreign workers and ethnic diversity in the workplace, especially among the high skilled, tend to foster innovation at the firm level and that lowering costs of communication and cross-cultural exchanges is key to transforming the ethnic diversity of the workforce into the firm's competitive advantage (Parrotta et al. 2014a and 2014b). International workers with better knowledge of the destination language or widely spoken languages facilitate the global interchange of skills and stimulate the overall economic performance.

Adserà: Princeton University

Pytliková: VSB - Technical University of Ostrava

REFERENCES

Bakker, D., Müller, A. Velupillai,V., Wichmann, S. Brown, C.H., Brown, P., Egorov, D., Mailhammer, R., Grant, A. and Holman, E.W. (2009). "Adding typology to lexicostatistics: a combined approach to language Classification." *Linguistic Typology,* vol. 13(1), pp. 169-181.

Bauer, T., Epstein, G. and Gang, I. (2005). "Enclaves, language, and the location choice of migrants." *Journal of Population Economics,* vol. 18(4), pp. 649-662.

Beine, M., Docquier, F. and Ozden, C. (2011). "Diasporas", *Journal of Development Economics,* vol. 95 (1), pp. 30-41.

Belot, M. and Ederveen, S. (2012). "Cultural and institutional barriers in migration between OECD countries", *Journal of Population Economics*, vol. 25(3), pp. 1077-1105.

Belot, M., and Hatton, T.J. (2012). "Skill selection and immigration in OECD countries", *Scandinavian Journal of Economic, vol. 114(4), pp. 681-730.*

Bialystok, E. and Martin, M.M. (2004). "Attention and inhibition in bilingual children: evidence from the dimensional change card sort task", *Developmental Science,* vol.7(3), pp. 325–339.

Bleakley, H. and Chin, A. (2004). "Language skills and earnings: evidence from childhood immigrants", *Review of Economics and Statistics,* vol. 84 (2), pp. 481-496.

Bleakley H. and Chin, A.. (2010). "Age at arrival, English proficiency, and social assimilation among US immigrants," *American Economic Journal: Applied Economics,* vol. 2(1), pp. 165-192.

Borjas, G. J. (1999). "Immigration and welfare magnets", *Journal of Labour Economics,* vol. 17 (4), pp. 607-637.

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

Chiquiar, D. and Hanson, G. (2005). “International migration, self-selection, and the distribution of wages: evidence from Mexico and the U.S.”, *Journal of Political Economy,* vol. 113 (2), pp. 239–281.

Chiswick, B. R., and Miller, P.W. (1995). “The endogeneity between language and earnings: international analyses”, *Journal of Labor Economics* vol. 13 (2), pp. 246-288.

Chiswick, B. R., and Miller, P.W. (2002). “Immigrant earnings: language skills, linguistic concentrations and the business cycle”, *Journal of Population Economics,* vol. 15(1), pp. 31-57.

Chiswick, B. R., and Miller, P.W. (2005). “Linguistic distance: a quantitative measure of the distance between English and other languages”, *Journal of Multilingual and Multicultural Development,* vol. 26(1), pp. 1-11.

Chiswick, B. R., and Miller, P.W. (2007). “Computer usage, destination language proficiency and the earnings of natives and immigrants”, *Review of the Economics of the Household,* vol. 5 (2), pp. 129-157.

Chiswick, B. R., and Miller, P.W. (2010). “Occupational language requirements and the value of English in the US labor market”, *Journal of Population Economics,* vol. 23(1), pp. 353–372.

Clark, X., Hatton, T.J. and Williamson, J.G. (2007). “Explaining U.S. immigration, 1971-1998”, *The Review of Economics and Statistics,* vol. 89(2), pp. 359-373.

Docquier, F., Lohest, O. and Marfouk, A. (2007). “Brain drain in developing countries”, *World Bank Economic Review,* vol. 21, pp. 193-218.

Docquier, F. and Marfouk, A. (2006). “Dataset”. In (C. Ozden and M. Schiff, eds). *International Migration, Remittances and Development,* New York: Palgrave Macmillan.

Docquier, F. and Rapoport, H. (2012). “Globalization, brain drain and development”, *Journal of Economic Literature,* vol. 50(3), pp. 681-730.

Dustmann, Ch. (1994). “Speaking fluency, writing fluency and earnings of migrants”, *Journal of Population Economics,* vol.7, pp. 133–56.

Dustmann, Ch., and van Soest, A. (2001). “Language fluency and earnings: estimation with misclassified language índicators”, *The Review of Economics and Statistics,* vol. 83 (4), pp. 663-674.

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

Dustmann, Ch., and Fabbri, F. (2003). “Language proficiency and labour market performance of immigrants in the UK”, *Economic Journal,* vol. 113, pp. 695-717.

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

European Commission. (2002). “Candidate countries Eurobarometer (CCB) http://europa.eu.int/comm/public_opinion/cceb_en.htm

European Commission. (2006). Special Eurobarometer on “Europeans and their languages”, http://ec.europa.eu/public_opinion/archives/ebs/ebs_243_sum_en.pdf (last accessed:1. February 2012).

Giulietti, C. and Wahba, J. (2013) “Welfare migration“. In (A. F. Constant and K. F. Zimmermann, eds.), *International handbook on the economics of migration*, Cheltenham, UK, and Northampton, USA: Edward Elgar.

Goodman, S. W. (2010a). *Naturalisation policies in Europe: exploring patterns of inclusion and exclusion*. Florence: EUI, EUDO Citizenship Observatory.

Goodman, S. W. (2010b). ‘Integration requirements for integration’s sake? Identifying, categorising and comparing civic integration policies’, *Journal of Ethnic and Migration Studies*, vol. 36(5), pp. 75372.

Grogger, J. and Hanson, G.H. (2011). “Income maximization and the selection and sorting of international migrants", *Journal of Development Economics,* vol. 95 (1), pp. 42-57.

Hatton, T. J. and Williamson, J.G. (2005). “What Fundamentals Drive World Migration?”, in (G. Borjas and J. Crisp, eds), *Poverty, international migration and asylum*, pp. 15-38, New York: Palgrave Macmillan Ltd.

Hatton, J. T. and Williamson G. J. (2011). ‘Are third world emigration forces abating? ’ *Word Development,* vol. 39(1), pp. 20-32.

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

Isphording, I. and Otten, S. (2011). “Linguistic distance and the language fluency of immigrants”, Ruhr Economic Papers No. 274.

Isphording, I. and Otten, S. (2013). ‘The costs of Babylon – linguistic distance in applied economics’, *Review of International Economics*, vol. 21(2), pp. 177–385.

Joppke, C. (2007) ‘Beyond national models: civic integration policies for immigrants in Western Europe’, *West European Politics*, vol. 30(1), pp. 122.

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.

Lewis, M. P. (2009). *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

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.

Massey, D. S., Arango, J., Graeme H., Kouaci, A., Pellegrino, A. and Taylor, E.J. (1993). “Theories of international migration: A review and appraisal.” *Population and Development Review,* vol.19 (3), pp. 431–466.

McFadden, D. (1974). “The measurement of urban travel demand”, *Journal of Public Economics,* vol.3, pp.: 303–328.

McKenzie, D. and Rapoport, H. (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.

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

OECD. (2011). Dataset. “International migration database.” http://stats.oecd.org/Index.aspx?DatasetCode=MIG (last accessed: 5 October 2011).

Ortega, F. and Peri, G. (2009). “The causes and effects of international migrations: Evidence from OECD countries 1980-2005”, National Bureau of Economic Research Working Paper No. 14833, NBER.

Parrotta, P., Pozzoli, D. and Pytlikova, M. (2014). “Does labour diversity affect firm productivity?” *European Economic Review,* vol. 66, pp. 144–179.

Parrotta, P., Pozzoli, D. and Pytlikova, M. (2014). "The nexus between labor diversity and firm's innovation." *Journal of Population Economics,* vol. 27(2), pp. 303-364.

Pedersen, P.J., Pytlikova, M. and Smith, N. (2008). “Selection and network effects – migration flows into OECD countries, 1990-2000”, *European Economic Review,* vol. 52(7), pp. 1160-1186.

Toomet, O. (2011). “Learn English, not the local language! Ethnic Russians in the Baltic states”, *American Economic Review,* vol. 101(3), pp. 526–31.

Simpson, N. B. and Sparber, Ch. (2013). “The short- and long-run determinants of unskilled immigration into U.S. states”, *Southern Economic Journal,* vol. 80(2), pp. 414-438.

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

Sjastaad, L. A. (1962). “The costs and returns of human migration”, *Journal of Political Economy,* vol.70 (5), pp. 80-93.

United Nations, Department of Economic and Social Affairs, Population Division. (2008). “United Nations Global Migration Database (UNGMD)”, http://esa.un.org/unmigration/ (last accessed: 29 August 2011).

Wadensjö, E. (2007). “Migration to Sweden from the new EU member states.” IZA Discussion Paper No. 3190, IZA Bonn.

Waldrauch, H. (2006). ‘Acquisition of nationality’, in (Baubock, R., Ersbøll, E., Groenendijk, K. and Waldrauch, H., eds), *Acquisition and loss of nationality*, Volume 1. Amsterdam: Amsterdam University Press, 12182.

Weil, P. (2001). ‘Access to citizenship: a comparison of twenty-five nationality laws’, in (Aleinikoff, T.A. and Klusmeyer, D., eds), *Citizenship Today: Global Perspectives and Practices*. Washington, DC: Carnegie Endowment for International Peace, 1735.

World Bank. (2011). Dataset. “Database of global bilateral migration.” http://data.worldbank.org/data-catalog/global-bilateral-migration-database (last accessed: 6 April 2011)

Zavodny, M. (1997). “Welfare and the locational choices of new immigrants.” *Economic Review – Federal Reserve Bank of Dallas;* 2Q: 2-10.

Table 1: Descriptive statistics, definitions and sources

VARIABLES	Definition	Source	Obs	Mean	Sd	Min	Max
Ln Emigration Rate	Ln(migration inflow from i to j per source population)	Own data collection, see Tables A1 and A3	100519	-5.02481	2.5311	-14.041	4.7055
Flowij	migration inflow from i to j	Own data collection, see Tables A1 and A3	102472	1077	6874	0	946167
Ln Stock of Migrants_t-1	Ln(foreign population stock from i in j per source population) t-1	Own data collection, , see Tables A2 and A4	77192	-3.1663	2.8741	-12.217	6.5749
Stockij	foreign population stock from i in j	Own data collection, , see Tables A2 and A4	82277	18922	169793	0	12100000
Linguistic Proximity	Linguistic Proximity index between i and j countries using their main official language.	Own calculation based on Ethnologue, see Data section	215040	.139397	.24967	0	1
Linguistic Proximity All	Linguistic Proximity index set at the maximum proximity between two countries using any of their official & main languages	Own calculation based on Ethnologue, see Data section	215040	.25353	.32499	0	1
Linguistic Proximity Major	Linguistic Proximity index between i and j countries using language spoken by majority	Own calculation based on Ethnologue, see Data section	215040	.08262	.19175	0	1
Dyen	Dyen Linguistic Proximity between i and j countries using their main official language based on the similarity of samples of words from each language	Dyen et al.(1992)	100608	414.3834	277.7419	110.6	1000
Dyen All	Dyen Linguistic Proximity set at the maximum proximity between two countries using any of their official &main languages	Dyen et al.(1992)	147264	490.5674	299.5442	112.8	1000
Dyen Major	Dyen Linguistic Proximity between i and j countries using language spoken by majority	Dyen et al.(1992)	66976	371.2698	260.8502	110.6	1000
Levenshtein	Levenshtein linguistic distance between i and j countries using their main official language	Max Planck Institute for Evolutionary Anthropology	208320	87.63829	23.59133	0	106.39
Levenshtein All	Levenshtein linguistic distance set at the maximum proximity between two countries using any of their official & main languages	Max Planck Institute for Evolutionary Anthropology	212160	78.29216	30.35295	0	106.39
Levenshtein Major	Levenshtein linguistic distance between i and j countries using language spoken by majority	Max Planck Institute for Evolutionary Anthropology	194880	91.60195	18.60136	0	106.4
Common Language	A dummy for common language between i and j countries	Own calculation based on Ethnologue, see Data section	215040	.0516369	.22129	0	1
Ln Destination GDPperCap_t-1	Ln GDP per capita, PPP (constant 2005 international $) in destination j, t-1	WDI, World Bank	196224	10.0149	.43263	8.6204	11.2134
Ln Origin GDPperCap_t-1	Ln GDP per capita, PPP (const 2005 international $) in origin i, t-1	WDI, World Bank	151770	8.5001	1.2706	4.9418	11.7221
Ln Origin GDPperCap_t-1 sq	Ln GDP per capita, PPP (const 2005 intern $) in origin i squared, t-1	WDI, World Bank	151770	73.8664	21.5499	24.4212	137.4071
Ln Public Expenditure	Ln Public social expenditure as a percentage of GDP in destination j, t-1	OECD SOCX Database	184576	2.8722	.4834	.5306	3.5752
Ln Destination UnemplRate_t-1	Ln Unemployment, total (% of total labor force) in destination j, t-1	WDI, World Bank	173152	1.8385	.5532	.4055	3.1739
InsubUj1	Unemployment rate in j, t-1, substituted by country average in case of missing values	WDI, World Bank and own calculation	208320	1.8538	.5436	.4055	3.1739
Ujmissing	A dummy indicating a missing value for unemployment rate in j	WDI, World Bank and own calculation	215040	.1948	.3960	0	1
Ln Origin UnemplRate_t-1	Ln Unemployment, total (% of total labor force) in origin i, t-1	WDI, World Bank	74460	1.9603	.7094	-1.6094	4.0860
InsubUi1	Unemployment rate in i, t-1, substituted by country average in case of missing values	WDI, World Bank and own calculation	174840	1.9796	.8117	-1.6094	4.0860
Uimissing	A dummy indicating a missing value for unemployment rate in i	WDI, World Bank and own calculation	215040	.6537	.4758	0	1
Ln Population Ratio, t-1	Ln Share of population in destination j per population in country i, t-1	WDI, World Bank	196740	8.4127	2.8426	-3.0912	17.2586
Ln Distance in km	Ln Distance between capitals of destination j and origin i in km	Own extension of CEPII	212128	8.5916	.8880	2.2741	9.8839
Neighboring Dummy	Dummy variable for neighboring countries	Own extension of CEPII	215040	.0183	.1341	0	1
Historical Past Dummy	Dummy variable for countries ever in colonial relationship	Own extension of Rose (2004)	215040	.0202	.1408	0	1
Ln Origin Political Rights_t-1	Ln of Freedom House Index – Political Rights in origin i	Freedom in the World Scores	167880	1.0700	.7461	0	1.9459
Ln Origin Civil Rights_t-1	Ln of Freedom House Index – Civil Liberties in origin i	Freedom in the World Scores	167880	1.1379	.6456	0	1.9459
Dominant Genetic Distance	Dominant genetic distance between plurality groups, current match	Spolaore and Waciarg (2009)	207360	933.9762	720.1979	0	2760
Weighted Genetic Distance	Weighted genetic distance, current match	Spolaore and Waciarg (2009)	184512	941.7764	651.6594	0	2777.695

Table 2. Language proximity and migration rates to 30 OECD destination countries from all world countries of origin for 1980-2010.

	OLS	OLS	OLS	FE	FE	FE	FE	**FE**	Betas	Poisson	FE	FE
VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Linguistic Proximity	3.271***	-	3.343***	0.774***	0.766***	0.960***	0.732***	**0.209*** **	0.020***	0.508***	0.222***	0.201***
	(0.147)		(0.215)	(0.123)	(0.123)	(0.146)	(0.123)	**(0.066)**		(0.121)	(0.084)	(0.065)
Common Language Dummy		2.929***	-0.095	-	-	-	-	-	-	-	-0.019	-
		(0.169)	(0.254)								(0.094)	
Ln Stock of Migrants_t-1	-	-	-	-	-	-	-	**0.669*** **	0.760***	0.693***	0.669***	0.659***
								(0.009)		(0.020)	(0.009)	(0.009)
Ln Destination GDPperCapPPPj_t-1	-	-	-	0.564***	0.570***	0.803***	0.916***	**1.723*** **	0.202***	1.608***	1.723***	1.633***
				(0.135)	(0.134)	(0.229)	(0.165)	**(0.132)**		(0.481)	(0.132)	(0.134)
Ln Origin GDPperCapPPPi_t-1	-	-	-	-0.268***	1.353***	1.087**	1.404***	**0.072**	0.037	0.224	0.071	0.199
				(0.051)	(0.286)	(0.547)	(0.315)	**(0.267)**		(0.695)	(0.267)	(0.277)
Ln Origin GDPperCapPPPit-1 squared	-	-	-	-	-0.102***	-0.050	-0.104***	**-0.011**	-0.097	-0.016	-0.011	-0.028*
					(0.017)	(0.031)	(0.019)	**(0.016)**		(0.045)	(0.016)	(0.017)
Ln Destination Public Social Expenditure_t-1	-	-	-	-	-	0.315**	0.384***	**0.576*** **	0.056***	-0.020	0.577***	0.566***
						(0.142)	(0.112)	**(0.101)**		(0.282)	(0.101)	(0.099)
Ln Destination UnemplRate_t-1	-	-	-	-	-	-0.064*	-0.078***	**-0.051** **	-0.010**	-0.141**	-0.051**	-0.027
						(0.036)	(0.028)	**(0.025)**		(0.068)	(0.025)	(0.026)
Ln Origin UnemplRate_t-1	-	-	-	-	-	0.124***	0.030	**0.054*** **	0.017***	0.191***	0.054***	0.049**
						(0.026)	(0.024)	**(0.021)**		(0.054)	(0.021)	(0.021)
lnTrade_t-1	-	-	-	-	-	-	-	**-**	-	-	-	0.122***
												(0.011)
Ln Population Ratio_t-1	-	-	-	1.218***	1.439***	1.168***	1.330***	**0.582*** **	0.550***	-0.081	0.583***	0.630***
				(0.101)	(0.112)	(0.172)	(0.119)	**(0.101)**		(0.288)	(0.101)	(0.102)
Ln Distance in km	-	-	-	-1.069***	-1.071***	-1.004***	-1.084***	**-0.390*** **	-0.145***	-0.377***	-0.390***	-0.257***
				(0.048)	(0.048)	(0.048)	(0.049)	**(0.030)**		(0.044)	(0.030)	(0.030)
Neighboring Dummy	-	-	-	0.230	0.227	0.193	0.134	**-0.198** **		-0.122	-0.198**	-0.226***
				(0.171)	(0.171)	(0.158)	(0.166)	**(0.082)**		(0.127)	(0.082)	(0.081)
Historical Past Dummy	-	-	-	1.942***	1.943***	1.427***	1.899***	**0.261*** **		0.075	0.266***	0.154*
				(0.181)	(0.181)	(0.195)	(0.184)	**(0.092)**		(0.100)	(0.098)	(0.088)
Dominant Genetic Distance	-	-	-	-0.0002**	-0.0002**	-0.00003	-0.0002*	**0.00003**	0.009	0.001***	0.00003	0.00002
				(0.000)	(0.000)	(0.000)	(0.000)	**(0.000)**		(0.000)	(0.000)	(0.000)
Ln Origin Freedom Political Rightsi_t-1	-	-	-	0.027	0.040	0.146***	0.022	**0.017**	0.005	0.094	0.017	0.025
				(0.027)	(0.027)	(0.037)	(0.029)	**(0.023)**		(0.101)	(0.023)	(0.023)
Ln Origin Freedom Civil Rightsi_t-1	-	-	-	-0.084**	-0.087***	-0.174***	-0.135***	**-0.074*** **	-0.019***	-0.063	-0.074***	-0.061**
				(0.034)	(0.033)	(0.041)	(0.036)	**(0.028)**		(0.070)	(0.028)	(0.028)
Dummies for Substituted unemployment	NO	NO	NO	NO	NO	NO	YES	**YES**	YES	YES	YES	YES
Destination & Origin FE	NO	NO	NO	YES	YES	YES	YES	**YES**	YES	YES	YES	YES
Constant	-5.574***	-5.257***	-5.580***	-6.772***	-14.830***	-18.742***	-18.621***	**-23.576*** **		-5.613	-23.577***	-25.040***
	(0.044)	(0.040)	(0.046)	(1.684)	(2.249)	(4.028)	(2.616)	**(2.167)**		(6.581)	(2.167)	(2.248)
Observations	100,519	100,519	100,519	84,874	84,874	41,621	74,797	**51,257**	51,257	51,257	51,257	50,561
Adjusted R-squared	0.111	0.076	0.111	0.756	0.756	0.774	0.764	**0.899**	0.899		0.899	0.902

Notes: OLS estimates with and without fixed effects. Dependent Variable: Ln(Emigration Rate). All models include year dummies. Robust standard errors clustered at the country-pair level in parentheses. Columns (7)-(12) use substituted unemployment rates whenever there is a missing information on unemployment rates. Substituted unemployment indicates that missing unemployment rates were set at the mean per country. Column (8) shows beta-coefficients of continuous variables in column (7) *** p<0.01, ** p<0.05, * p<0.1.

Table 3. Robustness checks: Alternative measures of linguistic proximity (Dyen and Levenshtein linguistic indexes and/or controls for multiple official and main languages) and migration rates to OECD countries.

	First Official Language			All Official and Main Languages			Major Language		
Ling. Proximity/Distance measured by:	Ling.Prox	Levenshtein	Dyen	Ling.Prox	Levenshtein	Dyen	Ling.Prox	Levenshtein	Dyen
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Linguistic Proximity	0.209***	-0.144*	0.203***	0.192***	-0.199***	0.333***	0.355***	-0.218**	0.225**
	(0.066)	(0.076)	(0.077)	(0.054)	(0.058)	(0.066)	(0.085)	(0.099)	(0.096)
Z-score	*[0.020]****	*[-0.013]**	*[0.022]****	*[0.024]****	*[-0.023]****	*[0.039]****	*[0.027]****	*[-0.016]***	*[0.023]***
Constant	-23.576***	-23.373***	-18.884***	-23.712***	-23.297***	-18.885***	-23.517***	-24.027***	-16.534***
	(2.167)	(2.201)	(3.136)	(2.171)	(2.174)	(2.495)	(2.167)	(2.231)	(3.996)
Observations	51,257	49,709	27,495	51,257	50,865	38,612	51,257	48,016	18,906
Adj. R2	0.899	0.899	0.900	0.899	0.899	0.904	0.899	0.898	0.902

Notes: Dependent Variable: Ln(Emigration Rate); numbers in brackets and italic show z-scores (beta coefficients). Controls included: stock of migrants, economic variables, distance variables, year dummies and destination and origin country fixed effects. Robust standard errors clustered at the country-pair level, *** p<0.01, ** p<0.05, * p<0.1.

Table 4. The role of English as widely spoken language, education and migration rates to OECD countries.

	All countries			Countries with low levels of education			All countries	
	First Official	Major	All Official & Main	First Official	Major	All Official & Main	First Official	First Official
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Linguistic Proximity:							0.244***	-0.014
							(0.067)	(0.126)
In Non-English destination	0.363***	0.509***	0.225***	0.271*	-0.176	0.368***		
	(0.073)	(0.082)	(0.059)	(0.144)	(0.287)	(0.099)		
In English destination	0.061	0.108	0.150*	0.025	0.108	0.227**		
	(0.095)	(0.147)	(0.083)	(0.123)	(0.237)	(0.100)		
Origin Tertiary Education_t							0.109***	0.099***
							(0.022)	(0.022)
Linguistic Prox*Ter Edu_t								0.094**
								(0.043)
Other controls	YES	YES	YES	YES	YES	YES	YES	YES
Constant	-23.579***	-23.579***	-23.686***	-41.942***	-41.629***	-41.945***	-23.650***	-23.725***
	(2.174)	(2.174)	(2.172)	(4.498)	(4.498)	(4.496)	(2.210)	(2.208)
Observations	51,257	51,257	51,257	11,079	11,079	11,079	50,497	50,497
Adj. R2	0.899	0.899	0.899	0.890	0.890	0.889	0.899	0.899

Notes: Dependent Variable: Ln(Emigration Rate). Linguistic Proximity measured by our Linguistic Proximity Index. A country with low education is below the 25th percentile in gross secondary school enrollment rates for a given year. Tertiary education is measured by gross enrollment rates. Controls included: stock of migrants, economic variables, distance variables, year dummies and destination and origin country fixed effects. Robust standard errors clustered at the country-pair level, *** p<0.01, ** p<0.05, * p<0.1.

Table 5. The role of policy, networks and linguistic networks on linguistic distance and migration rates to OECD countries.

	(1)	(2)	(3)	(4)
	Policy	Policy	Linguistic networks at the 4th level of the linguistic tree	Linguistic networks at the 3rd level of the linguistic tree
Linguistic Proximity	0.205***	0.244**	0.311***	0.467***
	(0.066)	(0.096)	(0.079)	(0.085)
Linguistic Requirement (Policy)_t	-0.249***	-0.240***		
	(0.027)	(0.031)		
Ling.Req.Policy_t *Ling. Prox		-0.065		
		(0.107)		
Linguistic networks_t-1			0.040***	0.027**
			(0.011)	(0.011)
Ling. Networks_t-1 *Ling. Prox			-0.035**	-0.065***
			(0.017)	(0.017)
Ln Stock of Migrants_t-1	0.671***	0.671***	0.655***	0.661***
	(0.009)	(0.009)	(0.010)	(0.010)
Constant	-23.374***	-23.374***	-23.847***	-23.770***
	(2.134)	(2.134)	(2.163)	(2.165)
Observations	51,233	51,233	51,147	51,112
Adj. R2	0.899	0.899	0.899	0.899

Notes: Dependent Variable: Ln(Emigration Rate). Linguistic Proximity measured by our Linguistic Proximity Index. Controls included: stock of migrants, economic variables, distance variables, year dummies and destination and origin country fixed effects. Linguistic networks measured on the fourth (column 1) and third (column 2) level of the linguistic family tree. Linguistic requirement for naturalization in destination countries is coded as1 formal test; 0.5 informal test and 0 none. Robust standard errors clustered at the country-pair level, *** p<0.01, ** p<0.05, * p<0.1.

Appendix:

Appendix Table A1: Country-year coverage migration flows

Columns: Destination countries

Rows: Year

Cell: numbers of source countries, for which we have observations on the number of migrants for particular year

Dest	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	FRA	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	MEX	NLD	NOR	NZL	POL	PRT	SVK	SE	TUR	USA
Year																														
2010	208	190		217	198	135	193	203	113	183				144	208	179	184			141		194	213	212	124	148	212	194		197
2009	205	190	184	214	194	141	193	203	113	183		26		139	209	178	188	201	58	141	128	198	202	212	123	150	212	192	200	198
2008	204	190	182	214	194	143	194	203	113	183	120	21		142	208	178	187	198	57	146	126	195	202	213	205	143	212	192	196	196
2007	206	190	93	214	194	147	193	203	113	183	124	19	191	128	2	178	181	197	28	142	126	197	202	213	205	126	211	192	195	197
2006	206	190	96	214	194	142	193	202	108	183	120	34	190	133	2	178	182	195	10	139		193	202	213	205	128	208	192	193	193
2005	203	190	85	214	194	142	191	203	66	183	107	114		121	2	178	185	10	10	137		187	202	213	205	124	208	192	193	195
2004	203	190	71	214	194	146	191	203	57	183	107	109		108	2	178	183	10	10	135		193	202	213	205	118	208	192	193	204
2003	201	189	70	214	195	142	191	203	57	183	127	107		121	2	178	180	10	10	127		191	202	213	205	114	208	192	194	204
2002	198	189	70	214	194	141	191	203	57	183	128	99		110	2	178	182	10	10	123		198	192	213	205	126	208	192	193	204
2001	198	189	70	214	194	115	84	203	57	183	130	106		117	2	178	181	10	10	116		197	192	213	205	114	208	192	194	204
2000	200	189	70	214	180	110	83	203	59	183	129	111		118	2	178	182	15	10	124		197	192	213	205	113	208	192	194	204
1999	198	189	70	214	180	108	193	203	58	183	118	110		114	2	178	181	15		123		191	192	213	205	114	208	159	172	204
1998	193	189	70	214	180	122	193	203	59	183	117	116	188	114	2	178	182	14		120		191	192	213	16	144	208	166	171	204
1997	192	189	55	214	179	111	193	203	39	183	118	48	183	114	2	178	179	14		110		194	192	213	14	144	208	164	172	204
1996	195	189	55	214	176	114	193	203	58	183	118	52	205	116	2	178	178	14		108		191	191	213	14	144	208	167	165	203
1995	187		55	214	176	117	193	203	39	183	118	54	203	117	2	178	48	15		110		187	192	213	13	144		165	165	203
1994	186		55	214	179	106	193	203	39	183	118	27	205	119	2	178	32	14		103		186	192	213	13	144		164		203
1993	180		48	214	178	97	193	203	39	183		39	205	106	2	178	32	14		99		185	192	213	11	143		168		204
1992	182		48	214	174		189	203	45	183		45	205	111	2	178	32	14		105		174	191	213	11	143		157		205
1991	171		48	213	158		172	203	42	183		49	206	104	2	178	32	11		95		160	191	213	11			148		206
1990	168		48	213	156		44	203	42	183		38	200	102	2	178	32	12		100		163	190	213	10			144		205
1989	155		48	213	154		105	203	42	183		31		97	2	178	32	11		93		164	192	213	10			142		205
1988	150		25	213	159		105	203	42	183		38		100	2	178	32	11		94		158	192	213				138		205
1987	159		27	213	155		105	203		183		29		99	2	178	32	7		93		161	192	213				136		205
1986	153		27	213	154		105	203		183		33		103		178	32	7					191	213				138		205
1985	155		27	213	154		105	203		183		35		95		18	32	7					116	213				134		205
1984	154		27	213	151		105	203		183						18							205	213				126		205
1983	166		27	213	152		105	203		183						18							205	213				123		205
1982	161		27	213	154		105	203								18							205	213				121		205
1981			27	213	154		105	203								18							205	213				123		204
1980			27	213			105	203															205	213				119		202
	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	FRA	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	MEX	NLD	NOR	NZL	POL	PRT	SVK	SE	TUR	USA

Appendix Table A2: Country-year coverage migration stocks

Columns: Destination Countries

Rows: Year

Cell: numbers of source countries, for which we have observations on the number of migrants for particular year

Dest	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	FRA	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR	LUX	MEX	MEX	NLD	NOR	NZL	POL	PRT	SVK	SE	TUR	USA
Year																															
2010		209			191	171	192	201		193		179		173	209	175	192			26			209	213		209	176	150	199		107
2009	209	209	185		194	172	190	201	112	191		171		180	208	175	190	201	27	26			207	213		209	177	145	199	185	133
2008	209	209	187		194	171	192	201	112	191	127	177		178		175	192	199	28	26			209	213			176	144	199	183	133
2007	209	209	178		194	168	193	200	112	191	128	174		174		175	188	198	25	26			207	213			179	142	199		133
2006	199	209	184	210	194	168	193	200	112	193	193	148	189	173	43	175	189	195	25	23			207	213	211		174	143	199		96
2005	209	209	182		194	166	139	201	112	193	204	97	191	165		175	189	183	25	23	10		208	213			173	138	199		96
2004	208	209	181		194	165	139	201	112	193		101	189	162		172	188	18	25	23		10	208	213			171	137	199		96
2003	208	209	181		194	163	138	201	112	193		100	190	156		172	188	18	25	23			207	213			167	149	199		96
2002	208	209	181		194	161	138	201	99	193		100		158	177	172	186	42	25	23			207	213		201	167	148	199		96
2001	190	207	181	190	194	163	138	201	99	193		97		154		172	187	42	19	12			206	213	199		166	142	199		96
2000	207	191	176		195	161	138	201	99	193		102	207	163		172	184	122	19	137	201		206	213			163	140	199	196	132
1999	206		174		195	164	138	201	99	193	162	87		163		172	185	42	19	12		202	204	213			157	136	111		96
1998	206		174		195	158	138	201	99	193		104		161		172	38	42	19	12			204	213			154	144	111		96
1997	204		55		195	152	138	201	99	193		100	189	159		172	189	42	19	12			204	212			151	144	111		96
1996	192		55	201	195	153	138	201	63	193		90	205	157	36	65	50	18	19	12			204	212	52		150	139	111		96
1995	202		55		195	150	138	201	58	193		85	205	146		65	50	37	19	12			200	212			150	140	111		96
1994	49		55		195	145	137	201	58	193		87	205			66	50	18	19	12			9	212			146		107		126
1993	49		48		195		137	201	58	193		87	205			66	50	18	19	12			9	212			139		104		126
1992	49		48		194		132	201	58	193		82	205			66	185	18	17	12			9	212			129		101		126
1991	168		48	180	194		117	201	58	193		70	205		2	43	184	16	15	12			9	212	51		125		98		126
1990	49	70	48		194		118	201	57	193	76		205			60		42	15	82			9	212			120		100	12	127
1989			48		194		118	201	57	134			204			60		12		8			9	212			121		98		125
1988					194		118	201	57	134			204			60		12	3	8			9	212			119		98		125
1987					194		118	201	57	131			204			60		12	4	8			9	212			118		97		125
1986	75			42	194		118	201	57	125			204		2	60		12	9	8			9	212	75		115		94		125
1985					194		118	201	57	124			204			60		42					9	212			109		95		125
1984					194		118	201		191			204			60		12					9	187			103		89		125
1983					194		118	201					204			60		12					9	187			100				125
1982					194		118	201					204			60		12						193			83		85		125
1981	81		47	42	194		118	201					204		2	59		12						189	75		98				125
1980		64			194		116	201					204					42		79				190			90		95		128
Dest	AUS	AUT	BEL	CAN	CHE	CZE	DEU	DNK	ESP	FIN	FRA	GBR	GRC	HUN	IRL	ISL	ITA	JPN	KOR				NLD	NOR	NZL	POL	PRT		SE	TUR	USA

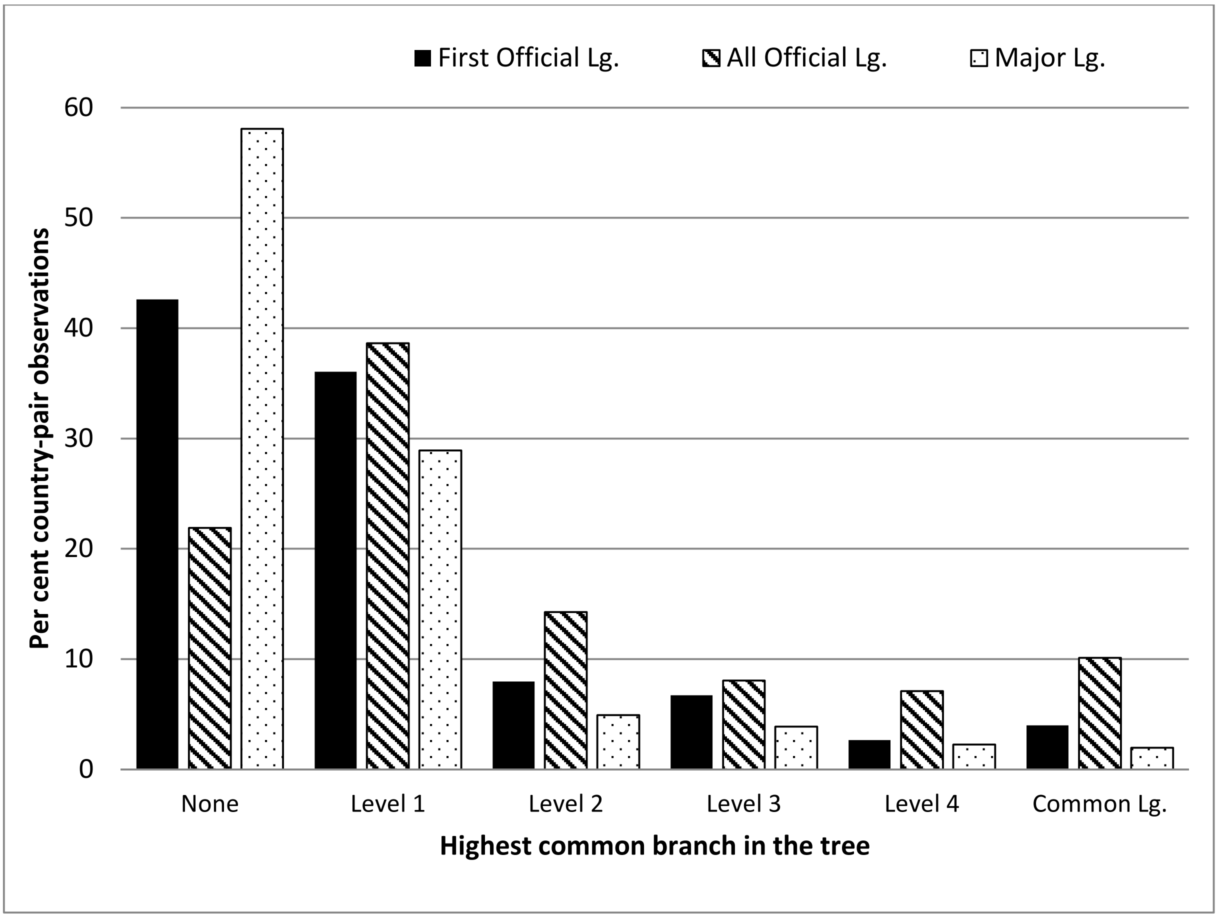
Appendix Table A3: Inflows of foreign population: definitions and sources

Migration flows to:	***Definition of "foreigner" based on***	***Source***
Australia	Country of Birth	Permanent and long term arrivals, Government of Australia, DIMA, Dept. of Immigration and Multicultural Affairs http://www.immi.gov.au/media/statistics/index.htm
Austria	Citizenship	Population register, Statistik Austria (1997 to 2002), Wanderungsstatistik 1996-2001, Vienna
Belgium	Citizenship	Population register. Institut National de Statistique.
Canada	Country of Birth	Issues of permanent residence permit. Statistics Canada – Citizenship and Immigration Statistics. *Flow is defined as a sum of foreign students, foreign workers and permanent residents.* http://www.cic.gc.ca/english/resources/statistics/facts2009/glossary.asp
Czech Rep.	Citizenship	Permanent residence permit and long-term visa, Population register, Czech Statistical Office
Denmark	Citizenship	Population register. Danmarks Statistics
Finland	Citizenship	Population register. Finish central statistical office
France	Citizenship	Statistics on long-term migration produced by the 'Institut national d'études démographiques (INED)' on the base on residence permit data (validity at least 1 year) transmitted by the Ministry of Interior.
Germany	Citizenship	Population register. Statistisches Bundesamt
Greece	Citizenship	Labour force survey. National Statistical Service of Greece 2006-2007 Eurostat
Hungary	Citizenship	Residence permits, National Hungary statistical office.
Iceland	Citizenship	Population register. Hagstofa Islands national statistical office.
Ireland	Country of Birth	Labour Force Survey. Central Statistical Office. Very aggregate, only very few individual origins.
Italy	Citizenship	Residence Permits. ISTAT
Japan	Citizenship	Years 1988-2005: Permanent and long-term permits. Register of Foreigners, Ministry of Justice, Office of Immigration. Years 2006-2008: Permanent and long-term permits. OECD Source International Migration data
Korea	Citizenship	OECD Source International Migration data
Luxembourg	Citizenship	Population register, Statistical Office Luxembourg
Mexico	Citizenship	OECD Source International Migration data
Netherlands	Country of Birth	Population register, CBS
New Zealand	Last Permanent Residence	Permanent and Long-term ARRIVALS (Annual – Dec) Census, Statistics New Zealand
Norway	1979-1984 Country of Origin 1985-2009 Citizenship	Population register, Statistics Norway
Poland	Country of Origin	Administrative systems (PESEL, POBYT), statistical surveys (LFS, EU-SILC, Population censuses). Central Statistical Office of Poland
Portugal	Citizenship	Residence Permit, Ministry of Interior.
Slovak rep.	Country of Origin	Permanent residence permit and long-term visa, Slovak Statistical Office
Spain	Country of Origin	Residence Permit, Ministry of Interior
Sweden	Citizenship	Population register, Statistics Sweden
Switzerland	Citizenship	Register of Foreigners, Federal Foreign Office of Switzerland
Turkey	Citizenship	OECD Source International Migration data
United Kingdom	Citizenship	Residence permits for at least 12 months. IPS - office for national statistics, and EUROSTAT
United States	Country of Birth	US Census Bureau Current Population Survey (CPS); U.S. Department of Homeland Security: *Yearbook of Immigration Statistics.* Persons obtaining Legal Permanent Resident Status by Region and Country of birth www.dhs.gov/ximgtn/statistics/publications/LPR06.shtm)

Appendix Table A4: Stock of foreign population: definitions and sources

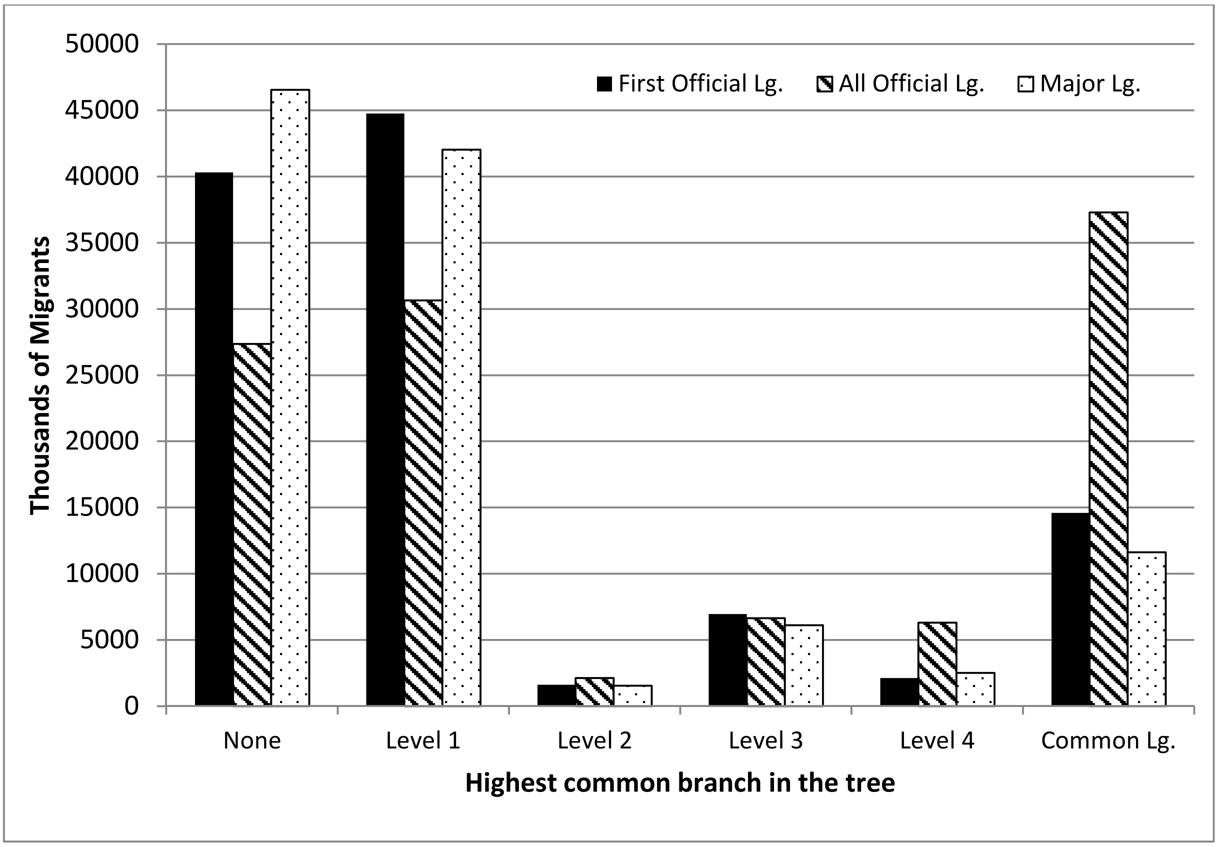
Foreign population stock in:	*Definition of "foreigner" based on*	*Source*
Australia	Country of birth	Census of Population and Housing, Australian Bureau of Statistics
Austria	Country of birth	Statistics Austria, Population Census 2001 and Population Register 2001 to 2009. For census year 1981 and 1991 definition by citizenship
Belgium	Citizenship	Population register. Institut National de Statistique
Canada	Country of birth	Census of Canada, Statistics Canada. www.statcan.ca/
Czech Rep.	Citizenship	Permanent residence permit and long-term visa, Population register, Czech Statistical Office and Directorate of Alien and Border Police
Denmark	Country of origin	Population register. Danmarks Statistics
Finland	Country of birth	Population register. Finish central statistical office
France	Country of birth	Census. Residence permit. Office des migrations internationals.
Germany	Citizenship	Population register. Statistisches Bundesamt
Greece	Citizenship	Labour force survey. National Statistical Service of Greece.
Hungary	Citizenship	National Hungary statistical office
Iceland	Country of birth	Population register. Hagstofa Islands
Ireland	Country of birth	Censuses, Statistical office, Ireland
Italy	Citizenship	Residence Permits. ISTAT
Japan	Citizenship	Years 1980-1999, Register of Foreigners, Ministry of Justice, Office of Immigration. Years 1999-2008 OECD Source Migration stat. Both sources based on permanent and long-term permits.
Korea	Citizenship	1986-1988: Trends in international migration Outlook, OECD 1990-2008: OECD Source International Migration Database
Luxembourg	Citizenship	Population register, Statistical office Luxembourg
Mexico	Country of birth	2005: Trends in international migration Outlook, OECD 2000: OECD Source International Migration Database
Netherlands	Citizenship	Population register, CBS
New Zealand	Country of birth	Census, Statistics New Zealand
Norway	Country background	Population register, Statistics Norway Country background is the person's own, their mother's or possibly their father's country of birth. Persons without an immigrant background only have Norway (000) as their country background. In cases where the parents have different countries of birth, the mother's country of birth is chosen.
Poland	Country of birth	2002 Census, rest permits, Statistics Poland
Portugal	Citizenship	Residence Permit, Ministry of Interior, www.ine.pt
Slovak Republic	Country of Origin	Permanent residence permit and long-term visa, Slovak Statistical Office
Spain	1985-1995 Citizenship 1996-2009 Country of birth	Residence Permit, Ministry of Interior
Sweden	Country of Birth	Population register, Statistics Sweden
Switzerland	Citizenship	Register of Foreigners, Federal Foreign Office
Turkey	Country of birth	OECD Source International Migration Database
United Kingdom	Country of Birth	LFS, UK statistical office
United States	Country of birth	US Census Bureau: 1990 and 2000 US census, the rest Current Population Survey (CPS) December. Data Ferret. Years 1980-1989, 1991-1993 from extrapolations by Tim Hatton (RESTAT)

Fig. 1: Distribution of country-pairs by linguistic proximity index



Notes: The linguistic index equals 0 if two languages do not belong to any common language family, 0.1 if they are only related at the level 1; 0.25 at level 2; 0.45 at level 3 and 0.7 at level 4. The index equals 1 if the two countries have a common language. The sample includes country-pairs in the baseline model in column (8) of Table 2. Unbalanced panel of 223 origin countries to 30 OECD destinations for period of 1980-2010.

Fig. 2: Distribution of migration flows by linguistic proximity index based on ethnolinguistic tree for years 1980-2010.



Notes: Migration flows are expressed in thousands of migrants. Unbalanced panel of 223 origin countries to 30 OECD destinations for period of 1980-2010.