# Linguistic Distance, Networks and the Regional Location Decisions of Migrants to the EU[*]

Julia Bredtmann[1,2], Klaus Nowotny[3,4], and Sebastian Otten[1,5]

[1] *Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)*
[2] *IZA*
[3] *University of Salzburg*
[4] *Austrian Institute of Economic Research WIFO*
[5] *CReAM/University College London*

Very preliminary version – please do not cite!

### Abstract

This paper analyzes the interaction between migrant networks and linguistic distance in the location decisions of migrants to the European Union at the regional level. We test the hypothesis that language and networks are substitutes in the location decision. Based on individual level data and a random utility maximization framework we find that networks have a positive effect on location decisions while the effect of linguistic distance is, as expected, negative. We also find a positive interaction effect between the two variables: networks are more important the larger the linguistic distance between the home and host countries, and the negative effect of linguistic distance is smaller the larger the network size.

*JEL Classifications:* F22, J61, R23

*Keywords:* Location choice, ethnic networks, linguistic distance, EU migration

[*] The authors are grateful to seminar participants at the ERSA congress 2015, the 2015 WIFO Regional Economics Workshop, the 2015 Jena Workshop on Migration Barriers, and the GfR winter seminar 2016 for helpful comments and suggestions. All correspondence to: Klaus Nowotny, University of Salzburg, Residenzplatz 9, A-5010 Salzburg, Austria, E-mail: klaus.nowotny@sbg.ac.at.

# 1 Introduction

Empirical evidence shows that migrant networks and diasporas are amongst the most important factors determining the location decisions of migrants (see, e.g., Gross and Schmitt, 2003; Pedersen *et al.*, 2008; Damm, 2009; Nowotny and Pennerstorfer, 2012; Beine *et al.*, 2011, 2015), even after controlling for income differences, employment opportunities, colonial ties and geographic distance. In addition, an emerging literature has identified language as another important factor for migrants' location decisions, either focusing on common language or, more recently, linguistic distance (see, e.g., Belot and Ederveen, 2012; Belot and Hatton, 2012; Adserà and Pytliková, 2015; Chiswick and Miller, 2015).

It can, however, be expected that the importance of networks depends on the linguistic distance and vice versa: networks can be expected to be more important the larger the linguistic distance between the home and host countries, and the negative effect of linguistic distance can be expected to be smaller for countries and regions with larger migrant networks. This argument is supported by McDonald (2004), who shows that French or English speaking immigrants to Canada are less likely to settle in an area with a large concentration of same ethnic group than immigrants who do not speak the host country's language.

This paper therefore contributes to the literature by analyzing the interaction between migrant networks and linguistic distance in the location decisions of migrants to the European Union. In doing so, we use individual level data from a special evaluation of the 2007 European Labour Force Survey (EU-LFS), which identifies migrants at the regional (NUTS-2) level, and a linguistic distance matrix based on the Levenshtein distance for a huge set of sending country-receiving region dyads. This enables us to capture within-country variation in linguistic distance and networks, respectively, and to analyze the location choice of migrants to the EU at the regional level.

Our results reveal that networks have a positive effect on the location decisions of migrants while the effect of linguistic distance is, as expected, negative. We also find a positive interaction effect between the two variables: networks are more important the larger the linguistic distance between the home country and the host region, and the negative effect of linguistic distance is smaller the larger the network size. This substitutable relationship between networks and language is robust to a number of sensitivity analyses.

The remainder of the paper is as follows. Section 2 outlines the empirical methodology and describes the data used. In Section 3, we show our estimation results and Section 4 concludes.

# 2 Method and Data

## 2.1 Method

The empirical analysis is based on a random utility maximization framework, in which migrant $i$ from sending country $s$ faces a set of alternative receiving regions $K$. The utility of the region $r \in K$ is represented by:

$$u_{isr} = \beta_1 \text{Network}_{sr} + \beta_2 \text{LD}_{sr} + \beta_3 \text{Network}_{sr} \times \text{LD}_{sr} + \gamma' X_{sr} + \varepsilon_{isr}, \tag{1}$$

where $Network_{sr}$ represents the size of the network of immigrants from source country $s$ in host region $r$ and $LD_{sr}$ denotes the linguistic distance between the source country and the host region. $X_{sr}$ is a set of control variables specific to source country $s$, region $r$, and the dyad $sr$, respectively, and $\varepsilon_{isr}$ is a random error term. The coefficient of main interest is $\beta_3$, the coefficient of the interaction between immigrant networks and linguistic distance. According to our hypothesis of a substitutable relationship between language and networks, we expect that $\beta_3 > 0$.

According to the behavioral model, migrant $i$ chooses region $r \in K$ if and only if $u_{isr} \geq u_{isk} \ \forall \ k \in K$. By assuming that the error term $\varepsilon_{isr}$ is i.i.d. extreme value, the probability that migrant $i$ chooses $r$ can be estimated by a conditional logit model (McFadden, 1974). Due to (largely) similar log-likelihood functions, we instead aggregate the data at the bilateral level and estimate the model using a Poisson pseudo-maximum likelihood estimator (PPML), as proposed by Guimarães *et al.* (2003), Santos Silva and Tenreyro (2006), and Schmidheiny and Brülhart (2011).

One problem associated with using PPML to estimate Eq. (1) is that it requires the observations to be cross-sectionally independent. If $X_{sr}$ fails to include all relevant bilateral determinants of migration or if some observed factors have an heterogeneous impact across potential migrants, then this would give rise to multilateral resistance to migration and the parameters in (1) would be exposed to an omitted variable bias (Bertoli and Fernández-Huertas Moraga, 2013, 2015). To address this problem, Bertoli and Fernández-Huertas Moraga (2015) suggest to add origin-nest fixed effects to Eq. (1) to control for unobservable nest-specific factors that have a differential impact on potential migrants from different countries of origin and this way restore the cross-sectional independence of the residuals in Eq. (1).

While this method has the advantage of being able to test the assumption of independence of error terms, it has two main disadvantages: First, the choice of nests is arbitrary and second, it requires to have enough variation in the data to identify the effect of interest after origin-nest fixed effects are included. The latter aspect is especially problematic in our case, as analyzing migration flows on a small regional level comes at the cost of having a higher number of zero observations, which raises multicollinearity issues. We therefore choose an alternative way to check whether multilateral resistance is a problem in our analysis.

[To be completed]

## 2.2 Data

The empirical analysis is based on individual level data from a special evaluation of the 2007 European Labour Force Survey (EU-LFS). The EU-LFS is a large household survey conducted each quarter among about 1.8 million persons aged 15 and above residing in the EU (see Eurostat, 2016, for an overview); annual data is also available and calculated from a combination of data collected on an annual and quarterly basis. While EU-LFS data disseminated by Eurostat usually contain only aggregated information on the sending countries, the microdata available to the authors provides detailed information on migrants' country of birth as well as their region of residence at the NUTS-2 level.

We define migrants as persons who were not born in their country of residence. As the data does not contain information on country of birth for Germany, we identify migrants to Germany based on nationality. For Ireland neither information on country of birth nor information on nationality is available, thus it has to be excluded from the analysis. The data further allow us to differentiate between those who moved to the EU between 1998 and 2007 and those who have been living in their host country for more than 10 years. The location choice will be modeled for migrants who moved to the EU-15 excluding Ireland (henceforth EU-14) between 1998 and 2007.[1] Overall, our data cover about 9 million migrants from 156 sending countries residing in 200 different receiving NUTS-2 regions.

One of our main explanatory variables is the migrant network in region $r$, which is defined as the stock of migrants from the same source country $s$ living in region $r$ in 2007 who migrated to country $C(r)$ before 1998. Following, amongst others, Beine *et al.* (2015) we assume a logarithmic form for network effect and add one to the network size in cases where it is zero, i.e.:

$$\text{Network}_{sr} = \ln(\text{Stock}_{sr}^{<1998} + 1).$$

Unfortunately, the EU-LFS does not allow a more detailed differentiation by year of arrival in the host country. Despite this shortcoming of the data, three arguments justify our approach: First, it takes some time for networks to be effective; only after previous

[1] Migration within the EU-15 is not considered because it is governed by a different migration regime than migration to the EU-15, which can affect the estimation results (see Razin and Wahba, 2015). Overseas territories as well as the Spanish exclaves Ceuta and Melilla are not considered as receiving regions. The same holds true for the relatively remote Canary Islands and the Azores and Madeira island regions. Moreover, due to its small population size, Denmark has to be considered as a single NUTS-2 region.

migrants have learned the administrative and social conventions of their host country, after they have found jobs or founded businesses providing ethnic goods, etc., they will be able to provide assistance to newly arrived members of their ethnic community. Second, by including only those who have been living in a region for at least 10 years our network variable includes only the most established members of a migrant's community. Although it could be argued that the tightness of links to the ethnic community decreases over time (for example, if previous migrants assimilate to the host country culture), these established members are likely to be the most helpful for newly arrived migrants. Third, because the network variable includes only those who migrated before 1998, the network size is not affected by those who migrated between 1998 and 2007 for which we model the location decision (Nowotny and Pennerstorfer, 2012).

Our second main variable of interest is the linguistic distance between the source country and the host region. As our measure of linguistic distance, we use the Levenshtein distance, which is based on the Automatic Similarity Judgement Program (ASJP) developed by the German Max Planck Institute for Evolutionary Anthropology.[2] The Levenshtein distance is calculated by comparing pairs of words having the same meaning in two different languages according to their pronunciation. The average similarity across a specific set of words is then taken as a measure for the linguistic distance between the languages (Bakker *et al.*, 2009). $\text{LD}_{sr}$ is thus defined as the average phonetic similarity between the most commonly spoken language in the source country and the most commonly spoken language in the receiving region.[3] The interaction between the size of the ethnic network and the linguistic distance, $\text{Network}_{sk} \times \text{LD}_{sk}$, then serves as our variable of main interest.

As further dyad-specific control variables, we include the natural logarithm of the geographic distance between the capital of the source country and the largest city in the host region. We further control for whether the source and the host country share or have ever shared a colonial relationship (Mayer and Zignago, 2011) and for whether they have a common official language that is spoken by at least 9% of the population (Melitz and Toubal, 2014). In addition, we include source-country fixed effects to control for origin-specific push factors (Ortega and Peri, 2013) and host-region fixed effects at the NUTS-2 level to control for destination-specific pull factors.[4]

[2] This measure was first applied to economics by Isphording and Otten (2014), who analyze the effect of linguistic distance on the language fluency of immigrants in the US and Germany.

[3] An example of the calculation of the linguistic distance for selected word pairs as well as the closest and furthest languages in our sample based on the Levenshtein distance are shown in Tables A1 and A2.

[4] Descriptive statistics of the control variables are shown in Table A3.

# 3 Results

## 3.1 Basic Results

Our main estimation results are shown in Table 1. We start with a specification that includes *Network* and *LD*, but no interaction between the two (Column I). In accordance with previous literature, we find that the size of the ethnic network has a positive effect on migrants' location choice (see, e.g., Beine *et al.*, 2011; Nowotny and Pennerstorfer, 2012), while the effect of linguistic distance is negative (see, e.g., Belot and Ederveen, 2012; Adserà and Pytliková, 2015). In column II, we add our variable of main interest, the interaction between networks and linguistic distance. We find a positive relationship between the interaction term and migrants' location decision, while sign and the significance of the single components of *Network* and *LD* remains stable. This supports the hypotheses that networks and language are substitutes in migrants' location choice: networks are more important the larger the linguistic distance between the home and host countries or, stated differently, the negative effect of linguistic distance is smaller the larger the network size. Importantly, the positive interaction effect between networks and linguistic distance remains after controlling for source- and destination-country fixed effects (Column III), further bilateral control variables, i.e., geographic distance and colonial relationship (Column IV), as well as after controlling for the existence of a common official language between the source and the host country (Column V).

The relationship between the other control variables and migrants' location choice is in line with previous literature. The geographic distance between the source country and the host region has a negative impact on the location choice. Moreover, as shown, amongst others, by Ortega and Peri (2009) and Grogger and Hanson (2011), people are more likely to migrate to countries that have a common colonial history. Lastly, migrants are attracted to countries that have a common official language, which considerably reduces migration costs (see Pedersen *et al.*, 2008) and can raise the returns-to-skill in the host country (Grogger and Hanson, 2011).

To get an idea of the magnitude of the interaction effect between linguistic distance and migrant networks, Table 2 shows the predicted probabilities of migrating from country $s$ to region $r$ after a one standard deviation increase in $Network_{sr}$ and $LD_{sr}$, respectively.[5] If *LD* equals zero, i.e., if the sending country and the receiving region share a common language, a one standard deviation increase in *Network* increases the probability of migrating to that region by 19 percent. At the $25^{th}$-percentile of the distribution of *LD*, however, a similar change in *Network* increases the odds of migrating by 46 percent, and at the maximum of the distribution of *LD*, a one standard deviation increase in *Network* is associated with a 51 percent increase in the probability to migrate. Similarly, the negative

[5] The results are based on the specification shown in Column V of Table 1.

effect of *LD* varies over the distribution of the *Network*: While a one standard deviation increase in *LD* is associated with a 22 percent decrease in the probability of migrating a the bottom end of the distribution of *Network*, i.e., when the network is zero, this negative effect decreases close to zero percent at the very top of the distribution of *Network*.

## 3.2 Sensitivity Analyses

To check the robustness of our results, we conduct several sensitivity analyses. The respective results are shown in Table 3. First, we include a measure for the genetic distance between the source and the host country as an additional control variable.[6] Genetic distance is usually used as a proxy for the cultural distance between countries and populations, respectively (see, e.g., Spolaore and Wacziarg, 2009), which should raise individual migration costs. As is evident from Column I, genetic distance has no explanatory power for the location choice of migrants to the EU, and the coefficients of the other covariates remain stable in both size and significance when genetic distance is controlled for. Hence, unobserved cultural differences between the source and the host country are not the main drivers of our results.

Second, we employ an alternative measures of the migrant network to check the robustness of our results. As argued by Nowotny and Pennerstorfer (2012), there is a large heterogeneity in the size of ethnic networks in Europe. A regional network of a given absolute size may be more important for new migrants coming from a small ethnic group than for those coming from a very large ethnic group, a heterogeneity that might affect our estimation results. *Relative Network* is therefore calculated as the stock of migrants from source country $s$ living in region $r$ for at least 10 years divided by the total number of migrants from that source country living in the EU for at least 10 years. The respective estimation results are shown in Column II of Table 3. As the absolute network, the relative network of past migrants is positively correlated with current migration flows. Moreover, the interaction effect between linguistic distance and the relative network is positive and highly significant, corroborating the hypothesis of networks and linguistic proximity to be substitutes in migrants' location choice.

As a third robustness check, we test whether our results remain stable when we use an alternative measure for linguistic distance or proximity between countries. Specifically, we interact our network variable with the indicator variable for whether the source and the host country share an official language. While this variable does not capture within-country variation in linguistic distance, it acknowledges the fact that migration costs might be lower for individuals that migrate between countries that share an official language, although

[6] The genetic distance measure as defined by Cavalli-Sforza *et al.* (1994) is related to the inverse probability that groups of alleles are the same for two populations. Hence, the lower the common frequency of alleles in two populations, the longer these populations have been separated.

the most common spoken language is not the same. As can be seen from Column III of Table 3, our basic interpretation remains the same when using this alternative measure for linguistic proximity. The interaction effect between networks and sharing an official language is negative and significant, suggesting that the positive effect of networks on migration flows is smaller for countries with a high linguistic proximity.

Lastly, we restrict our sample to observations with positive values of $LD$, i.e., we eliminate migration flows between source countries and host regions that have the same most common language. While these observations represent only a small proportion of our overall sample (2.4 percent), we still want to rule out that the large migration flows between regions that have the same language are the main drivers of our results. The respective estimation results are shown in Column IV of Table 3. When excluding observations with $LD = 0$, the coefficient of the single component of the network effect turns negative (significant only at 10-percent level). The interaction effect between linguistic distance and networks remains positive and significant, and largely increases in magnitude. This suggests that the positive effect of ethnic network only comes into play for higher levels of linguistic distance. However, given that $LD > 0$ is effectively bound between 39.6 and 105.4 (see Table A2), the effect of network is *de facto* positive for the sample considered. This becomes obvious from Figure 1, which shows the estimated elasticity between networks and migrations flows over the range of the strictly positive $LD$ measure. The network effect increases from zero at the very bottom of the $LD$ distribution to about 0.32 at the very top of the $LD$ distribution, suggesting that at maximum levels of $LD$, a one percent increase in the network size increases bilateral migration flows by 0.32 percent.

# 4 Conclusion

In this paper, we investigate the role of language and networks in the regional location choice of migrants to the EU. In particular, we are interested in whether regional ethnic networks and linguistic proximity represent substitutes in migrants' location decision. Our empirical analysis is based on a random utility maximization framework and employs individual level data from a special evaluation of the 2007 European Labor Force Survey, which allows us to identify migrants at the regional (NUTS-2) level.

Our results reveal that both ethnic networks and linguistic distance are important determinants of the regional location choice of migrants to the EU. Furthermore, networks and linguistic proximity represent substitutes in migrants' location choice: Regional ethnic networks become more important the higher linguistic distance between the source country and the receiving region, and the negative effect of linguistic distance decreases with increasing network size. These results are robust to a number of sensitivity analyses.

Although we are not able to identify a causal impact of networks on migrants' location choice, because ethnic networks themselves might be affected by a number of different factors, including linguistic proximity, our results have some important implications for the possible direction of future migration flows. In the next years, huge flows of refugees are expected to enter Europe. Given that the situation in their countries of origin does not change substantially, many are going to settle in their new destinations permanently. Our findings suggest that such newly established networks are able to substantially reduce linguistic barriers and this way shape future migration flows.

# References

Adserà, A. and Pytliková, M. (2015). The role of language in shaping international migration: Evidence from OECD countries 1985-2006. *The Economic Journal*, **125** (586), F49–F81.

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*, **13** (1), 169–181.

Beine, M., Docquier, F. and Özden, a. (2011). Diasporas. *Journal of Development Economics*, **95** (1), 30–41.

—, — and Özden, a. (2015). Dissecting Network Externalities in International Migration. *Journal of Demographic Economics*, **81**, 379–408.

Belot, M. V. K. and Ederveen, S. (2012). Cultural barriers in migration between OECD countries. *Journal of Population Economics*, **25** (3), 1077–1105.

— and Hatton, T. J. (2012). Immigrant Selection in the OECD. *The Scandinavian Journal of Economics*, **114** (4), 1105–1128.

Bertoli, S. and Fernández-Huertas Moraga, J. (2013). Multilateral resistance to migration. *Journal of Development Economics*, **102**, 79–100.

— and Fernández-Huertas Moraga, J. (2015). The size of the cliff at the border. *Regional Science and Urban Economics*, **51**, 1–6.

Cavalli-Sforza, L. L., Menozzi, P. and Piazza, A. (1994). *The history and geography of human genes*. Princeton: Princeton University Press.

Chiswick, B. R. and Miller, P. W. (2015). International Migration and the Economics of Language. In B. R. Chiswick and P. W. Miller (eds.), *Handbook of the Economics of International Migration 1A*, Oxford and Amsterdam: North-Holland, pp. 211–269.

Damm, A. P. (2009). Determinants of recent immigrants' location choices: quasi-experimental evidence. *Journal of Population Economics*, **22** (1), 145–174.

Eurostat (2016). Statistics Explained – EU labour force survey. http://ec.europa.eu/eurostat/statistics-explained/index.php/EU_labour_force_survey, accessed 16/04/04.

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

Gross, D. M. and Schmitt, N. (2003). The Role of Cultural Clustering in Attracting New Immigrants. *Journal of Regional Science*, **43** (2), 295–318.

Guimarães, P., Figueirdo, O. and Woodward, D. (2003). A Tractable Approach to the Firm Location Decision Problem. *The Review of Economics and Statistics*, **85** (2), 201–204.

Isphording, I. E. and Otten, S. (2014). Linguistic Distance and the Language Fluency of Immigrants. *Journal of Economic Behavior & Organization*, **105**, 30–50.

Mayer, T. and Zignago, S. (2011). Notes on CEPII's distances measures: The GeoDist database. CEPII Working Paper 2011-25.

McDonald, J. (2004). Ethnic clustering and the location choice of immigrants to Canada. *Canadian Journal of Urban Research*, **13** (1), 85–101.

McFadden, D. (1974). Conditional logit analysis of qualitative choices. In P. Zarembka (ed.), *Frontiers in Econometrics*, New York: Academic Press, pp. 105–142.

Melitz, J. and Toubal, F. (2014). Native language, spoken language, translation and trade. *Journal of International Economics*, **93** (2), 351–363.

Nowotny, K. and Pennerstorfer, D. (2012). Ethnic Networks and the Location Choice of Migrants in Europe. University of Salzburg Working Paper in Economics and Finance No. 2012-07.

Ortega, F. and Peri, G. (2009). The Causes and Effects of International Migrations: Evidence from OECD Countries 1980-2005. NBER Working Paper No. 14833.

— and — (2013). The effect of income and immigration policies on international migration. *Migration Studies*, **1** (1), 47–74.

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

Razin, A. and Wahba, J. (2015). Welfare Magnet Hypothesis, Fiscal Burden, and Immigration Skill Selectivity. *The Scandinavian Journal of Economics*, **117** (2), 369–402.

Santos Silva, J. M. C. and Tenreyro, S. (2006). The Log of Gravity. *The Review of Economics and Statistics*, **88** (74), 641–658.

Schmidheiny, K. and Brülhart, M. (2011). On the equivalence of location choice models: Conditional logit, nested logit and Poisson. *Journal of Urban Economics*, **69** (2), 214–222.

Spolaore, E. and Wacziarg, R. (2009). The Diffusion of Development. *The Quarterly Journal of Economics*, **124** (2), 469–529.

# Figures

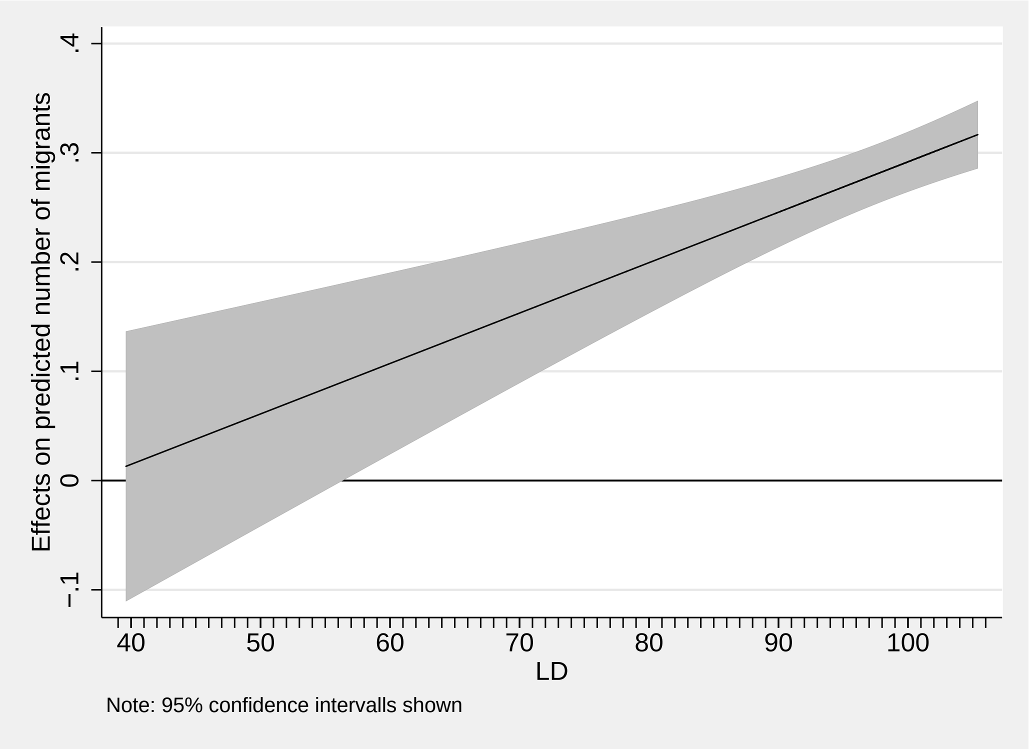

**Figure 1:** Effect of Network over LD $> 0$ (in Elasticities)

# Tables

**Table 1:** PPML Estimation of Migration Flows to the EU

| | I<br>Coef/StdE | II<br>Coef/StdE | III<br>Coef/StdE | IV<br>Coef/StdE | V<br>Coef/StdE |
|---|---|---|---|---|---|
| Network | 0.4985$^{\dagger}$ | 0.3780$^{\dagger}$ | 0.1300$^{\dagger}$ | 0.1151$^{\dagger}$ | 0.1114$^{\dagger}$ |
| | (0.0196) | (0.0640) | (0.0295) | (0.0269) | (0.0265) |
| LD | −0.0118$^{\dagger}$ | −0.0221$^{\dagger}$ | −0.0334$^{\dagger}$ | −0.0231$^{\dagger}$ | −0.0191$^{\dagger}$ |
| | (0.0013) | (0.0041) | (0.0029) | (0.0032) | (0.0033) |
| Network × LD | – | 0.0014$^{**}$ | 0.0019$^{\dagger}$ | 0.0015$^{\dagger}$ | 0.0014$^{\dagger}$ |
| | | (0.0007) | (0.0003) | (0.0003) | (0.0003) |
| ln(distance) | – | – | – | −0.3823$^{***}$ | −0.4592$^{\dagger}$ |
| | | | | (0.1263) | (0.1258) |
| Colony | – | – | – | 0.9714$^{\dagger}$ | 0.4476$^{\dagger}$ |
| | | | | (0.1138) | (0.1112) |
| Common off. lang. | – | – | – | – | 0.9558$^{\dagger}$ |
| | | | | | (0.1267) |
| Constant | −2.2992$^{\dagger}$ | −1.4275$^{\dagger}$ | −5.3086$^{\dagger}$ | −3.4647$^{***}$ | −3.8880$^{***}$ |
| | (0.1835) | (0.4068) | (1.0986) | (1.3054) | (1.3140) |
| Fixed effects | no | no | yes | yes | yes |
| R$^2$ | 0.203 | 0.204 | 0.671 | 0.703 | 0.708 |
| Observations | 31,194 | 31,194 | 31,194 | 31,194 | 31,194 |

*Notes:* – $^{\dagger}$ $p < 0.001$; $^{***}$ $p < 0.01$; $^{**}$ $p < 0.05$; $^{*}$ $p < 0.1$. *– Robust standard errors in parentheses. – LD: linguistic distance. – PPML: Poisson pseudo-maximum-likelihood.*

**Table 2:** Change in Predicted Probabilities after a One SD Increase in Network and LD

| | Network | | |
|---|---|---|---|
| LD | Median | Median+SD | % change |
| Zero | 0.142 | 0.169 | 19.4 |
| P10 | 0.092 | 0.129 | 40.0 |
| P25 | 0.082 | 0.120 | 46.4 |
| P50 | 0.079 | 0.118 | 48.1 |
| P75 | 0.078 | 0.116 | 49.3 |
| P90 | 0.077 | 0.115 | 49.9 |
| Max | 0.075 | 0.113 | 51.4 |

| | LD | | |
|---|---|---|---|
| Network | Median | Median+SD | % change |
| Zero | 0.009 | 0.007 | −22.0 |
| P10 | 0.046 | 0.041 | −11.7 |
| P25 | 0.060 | 0.054 | −9.9 |
| P50 | 0.079 | 0.073 | −7.9 |
| P75 | 0.102 | 0.096 | −6.1 |
| P90 | 0.131 | 0.125 | −4.3 |
| Max | 0.205 | 0.203 | −0.9 |

*Notes: – All other variables at mean values. – P10, P25, etc. refer to the 10th, 25th, etc. percentile of the distribution of nonzero Network and LD. "Median" and "Median+SD" refer to the median and the median plus one standard deviation of the nonzero Network and LD.*

**Table 3:** PPML Estimation of Migration Flows: Robustness Checks

| | I<br>Coef/StdE | II<br>Coef/StdE | III<br>Coef/StdE | IV<br>Coef/StdE |
|---|---|---|---|---|
| Network | $0.1109^{\dagger}$ | – | $0.2424^{\dagger}$ | $-0.1698$ |
| | (0.0265) | | (0.0190) | (0.1033) |
| LD | $-0.0192^{\dagger}$ | $-0.0143^{\dagger}$ | $-0.0082^{\dagger}$ | $-0.0656^{\dagger}$ |
| | (0.0033) | (0.0024) | (0.0022) | (0.0071) |
| Network × LD | $0.0014^{\dagger}$ | – | – | $0.0045^{\dagger}$ |
| | (0.0003) | | | (0.0011) |
| Relative network | – | $0.0205^{***}$ | – | – |
| | | (0.0066) | | |
| Relative network × LD | – | $0.0006^{\dagger}$ | – | – |
| | | (0.0001) | | |
| Network × Common off. lang. | – | – | $-0.0846^{\dagger}$ | – |
| | | | (0.0254) | |
| ln(distance) | $-0.4587^{\dagger}$ | $-0.8535^{\dagger}$ | $-0.4273^{\dagger}$ | $-0.4658^{\dagger}$ |
| | (0.1270) | (0.1385) | (0.1197) | (0.1293) |
| Colony | $0.4438^{\dagger}$ | $0.5711^{\dagger}$ | $0.4995^{\dagger}$ | $0.4631^{\dagger}$ |
| | (0.1113) | (0.1075) | (0.1084) | (0.1105) |
| Common off. lang. | $0.9609^{\dagger}$ | $1.1553^{\dagger}$ | $1.6144^{\dagger}$ | $1.0273^{\dagger}$ |
| | (0.1268) | (0.1308) | (0.2251) | (0.1362) |
| Genetic distance | $-0.0210$ | – | – | – |
| | (0.2940) | | | |
| Constant | $-1.5408$ | $-1.9284$ | $-4.9503^{\dagger}$ | 0.0888 |
| | (1.2467) | (1.3655) | (1.3120) | (1.3551) |
| Fixed effects | yes | yes | yes | yes |
| Sample LD = 0 incl. | yes | yes | yes | no |
| $R^2$ | 0.708 | 0.649 | 0.707 | 0.675 |
| Observations | 30,794 | 31,194 | 31,194 | 30,451 |

*Notes:* – $^{\dagger}$ $p < 0.001$; $^{***}$ $p < 0.01$; $^{**}$ $p < 0.05$; $^{*}$ $p < 0.1$. *– Robust standard errors in parentheses. – LD: linguistic distance. – PPML: Poisson pseudo-maximum-likelihood. – Information on genetic distance is not available for Andorra an the State of Palestine, reducing the sample by 400 observations. – Relative Network is calculated as the stock of migrants from source country s living in region r divided by the total number of migrants from that source country living in the EU, i.e., Relative Network* $= (stock_{sr}^{\leq 1998}/stockEU_{s}^{\leq 1998}) \times 100$.

# Appendix

**Table A1:** Example: Computation of Word Distance

| **Word** | **English** | **German** | **Minimum Distance** |
|---|---|---|---|
| fish | *fiS* | fiS | 0 |
| breast | *brest* | brust | 1 |
| hand | *hEnd* | hant | 2 |
| tree | *tri* | baum | 4 |
| Mountain | *maunt3n* | bErk | 7 |

*Notes: – Averaged and normalized to account for differences in word length and similarities by chance.*

**Table A2:** Closest and Furthest Language Pairs in the Sample

| Closest | | Furthest | |
|---|---|---|---|
| Language | Distance | Language | Distance |
| *Distance to English* | | | |
| Jamaican Creole | 39.61 | Sar Chad (Chad) | 102.50 |
| Tok Pisin (Papua New Guinea) | 51.99 | Somali (Somalia) | 102.86 |
| Dutch | 60.73 | Fulfulde Adamawa (Guinea) | 103.10 |
| Norwegian | 61.41 | Vietnamese | 103.81 |
| Swiss German | 71.29 | Turkmen (Turkmenistan) | 104.54 |
| *Source-country and host-region language pairs* | | | |
| English<br>Jamaican Creole | 39.61 | Catalan<br>Swahili (Tanzania) | 105.13 |
| Finnish<br>Estonian | 47.55 | Danish<br>Palestinian Arabic | 105.27 |
| Danish<br>Norwegian | 47.85 | Greek<br>Swazi (Swaziland) | 105.39 |

*Notes: – The table shows the five closest and furthest languages toward English and the three closest and furthest source-country and host-region language pairs according to the normalized and divided Levenshtein distance. – Only languages spoken within the estimation sample are listed. – Geographic origin of language in parentheses.*

**Table A3:** Descriptive Statistics

| | Mean | StdD | Min | Max |
|---|---|---|---|---|
| Network | 1.020 | 2.430 | 0.000 | 12.491 |
| LD | 92.085 | 17.445 | 0.000 | 105.390 |
| Network × LD | 87.420 | 219.399 | 0.000 | 1, 215.394 |
| ln(distance) | 8.487 | 0.760 | 4.009 | 9.900 |
| Colony | 0.101 | 0.302 | 0.000 | 1.000 |
| Common off. lang. | 0.093 | 0.291 | 0.000 | 1.000 |
| Observations | | 31,194 | | |