CReAM
Centre for Research & Analysis of Migration

Discussion Paper Series

CPD 12/17

Linguistic Distance, Networks and Migrants' Regional Location Choice

Julia Bredtmann, Klaus Nowotny, and Sebastian Otten

Centre for Research and Analysis of Migration
Department of Economics, University College London
Drayton House, 30 Gordon Street, London WC1H 0AX

www.cream-migration.org

Linguistic Distance, Networks and Migrants' Regional Location Choice*

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

[1] *RWI, IZA, CReAM*
[2] *University of Salzburg, Austrian Institute of Economic Research WIFO*
[3] *University College London, CReAM, RWI*

Abstract

This paper analyzes the interaction between migrant networks and linguistic distance in the location choice of migrants to the EU at the regional level. We test the hypothesis that networks and the ability to communicate in the host country language, proxied by linguistic distance, are substitutes in the location decision. Based on individual level data from a special evaluation of the European Labour Force Survey (EU-LFS) and a random utility maximization framework, we find that networks have a positive effect on the location decisions while the effect of linguistic distance is negative. We also find a strong positive interaction effect between the two factors: 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. In several extensions and robustness checks, we show that this substitutable relationship is extremely robust.

JEL Classifications: F22, J61, R23

Keywords: Location choice, ethnic networks, linguistic distance, EU migration, multilateral resistance

*The authors are grateful to participants at the ERSA congress 2015, the 2015 WIFO Regional Economics Workshop, the 2015 Jena Workshop on Migration Barriers, the 13th IZA Annual Migration Meeting (AM^2), the GfR Winter Seminar 2016, and the 2017 AEA Meetings as well as to seminar participants at RWI, the University of Salzburg, and the Université catholique de Louvain 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

Previous research has shown that migrants tend to settle where other migrants of the same ethnicity or from the same country of birth have settled before. The importance of migrant networks and diasporas for the location choice of migrants even persists after taking into account other factors such as income differences, employment opportunities, colonial ties, and geographic distance (see, e.g., Pedersen *et al.*, 2008; Damm, 2009; Beine *et al.*, 2011, 2015; Nowotny and Pennerstorfer, 2017). The literature has identified a number of channels through which networks increase the attractiveness of a region for newly arriving immigrants. For instance, networks offer ethnic goods such as food, clothing, social organizations, religious services, media or marriage markets (Chiswick and Miller, 2005). Furthermore, established network members can provide information on housing or employment opportunities (Gross and Schmitt, 2003), assist with the settlement process or reduce legal entry barriers via family reunification programs (Beine *et al.*, 2015).

In addition, studies have shown that skills in the host country's language are an important determinant of migrants' labor market outcomes such as earnings and employment (see, e.g., Chiswick and Miller, 1995; Dustmann and Van Soest, 2002; Bleakley and Chin, 2004). The ability to communicate in the host-country language may affect the marginal productivity, enhance the capability to accumulate human capital, and affect the occupational choice of migrants (Chiswick and Miller, 2010; Isphording *et al.*, 2014). It may also affect other social outcomes with important economic consequences, as for example the probability of criminal activity or an individual's health status (Clarke and Isphording, 2017). Therefore, existing skills and the potential difficulties to acquire knowledge in the host country's language, as measured by the linguistic proximity between the source and the host country, are important factors for the location choice of migrants (see, e.g., Belot and Ederveen, 2012; Belot and Hatton, 2012; Adserà and Pytliková, 2015; Chiswick and Miller, 2015).

Whereas the effects of networks and linguistic proximity on migrants' location decision have been studied extensively, hardly anything is known on how the two effects relate to each other. From a theoretical perspective, we would expect the importance of networks to depend on the extent of language dissimilarities and vice versa (Lazear, 1999; Bauer *et al.*, 2005): networks should be more important the larger the dissimilarity between the languages of the home country and the host region, and the negative effect of linguistic dissimilarities should be smaller the larger the migrant network. Building on previous literature, the present paper therefore analyzes the interaction between migrant networks and linguistic proximity in the location decisions of migrants. The linguistic distance between the language of the home country and the host region is used as a measure of the proximity of the two languages, thereby indicating the difficulties migrants face in acquiring skills in the host region language (see Isphording and Otten, 2014). Analyzing the interaction between networks and linguistic distance improves our understanding of the factors that affect migration decisions and the location choice. The decision to migrate is an investment and a mechanism through which people try to improve their economic situation. However, the move to a different location, where the returns to skills are potentially higher, is costly. Knowledge about the potential substitutability of factors that reduce the migration costs or increase the economic returns is crucial to facilitate a more efficient allocation of productive resources.

We further contribute to the existing literature by analyzing migrants' location choice at a disaggregated regional level, thereby taking the sizeable regional differences within host countries into account. Only few studies have analyzed the location decision of international migrants from a regional perspective (see Åslund, 2005; Damm, 2009; Rodríguez-Pose and Ketterer, 2012; Nowotny and Pennerstorfer, 2017), hence within-country differences are largely underexplored in the migration choice literature. Yet, not only may the language spoken by the majority of the resident population differ between regions within a country, but migrant networks are also not equally distributed within a country. Besides regional variation in economic conditions that determine the attractiveness of a location in general, within-country differences in networks and language may provide an explanation why migrants (from a certain source country) cluster in some regions and not in others.

Lastly, our paper contributes to the literature by using a random parameters (mixed) logit framework (see McFadden and Train, 2000) to model migrants' location choices as an alternative way to deal with the issue of multilateral resistance to migration (Bertoli and Fernández-Huertas Moraga, 2013; Nowotny and Pennerstorfer, 2017).[1] By using this model we are able to relax the independence of irrelevant alternatives (IID) assumption inherent to the conditional logit model and its Poisson equivalent usually applied in empirical research (see, e.g., Guimarães *et al.*, 2003; Schmidheiny and Brülhart, 2011; Bertoli and Fernández-Huertas Moraga, 2015).

Our empirical analysis is based on individual level data from a special evaluation of the 2007 European Labour Force Survey (EU-LFS), which includes information on individuals' country of birth as well as their region of residence (at the NUTS-2 level). The data can be linked to a linguistic distance matrix based on the Levenshtein distance for a comprehensive 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 a very disaggregated regional level.

Our results reveal that networks have a significant and positive effect on the location decisions of migrants while the effect of linguistic distance is, as expected, negative. We also find a strong positive interaction effect between the two factors: 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. In several sensitivity analyses and extensions, we show that this substitutable relationship between networks and linguistic distance is extremely robust. Especially, we show that our results are not biased by multilateral resistance to migration. Taken together, these findings are consistent with our expectations that higher migration costs, due to higher language acquisition costs or a smaller network, can be offset by a larger network or by a lower linguistic distance.

[1] As outlined by Bertoli and Fernández-Huertas Moraga (2013), the rate of migration observed between two countries or regions does not depend solely on their relative attractiveness, but also on the one of alternative destinations. The term multilateral resistance to migration describes the influence exerted by other destinations on bilateral migration flows.

These results have interesting implications. First, a better knowledge of the relationship between two of the main determinants of migrants' location decision enhances our understanding of the location choice process. Hence, our findings have important implications for studies that investigate migration flows. Second, our results reveal that larger networks can decrease adjustment costs and increase the propensity to migrate to locations that are, from a linguistically perspective, very different to the migrant's home country. This suggests that large inflows of migrants from linguistically different origin countries, such as the current refugee influx to Europe, can substantially reduce the adverse effects of linguistic barriers for new migrants and this way shape future migration flows.

The remainder of the paper is as follows. Section 2 outlines the empirical methodology and describes the data used. In Section 3, we discuss our results and robustness analyses. Section 4 provides concluding remarks.

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:

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

where V_{isr} represents the deterministic component of utility and ε_{isr} is a random error term. V_{isr} is a function of the size of the network of immigrants from sending country s in receiving region r ($Network_{sr}$), the linguistic distance between the sending country and the receiving region (LD_{sr}), the interaction between the two ($Network_{sr} \times LD_{sr}$) as well as a set of further control variables (X_{sr}) specific to sending country s, receiving region r, and the dyad sr, respectively. The coefficient of main interest is β_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 $\beta_3 > 0$.

Deriving from 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 ε_{isr} is i.i.d. extreme value, the probability that migrant i chooses region 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 a 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.[2] The latter aspect is especially problematic in our context, 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.

[2] Furthermore, Bertoli and Fernández-Huertas Moraga (2015) apply a sequential approach where they (i) estimate the model with m nests, (ii) test for cross-sectional independence, and (iii) increase the number of nests m by one if the null hypothesis is rejected. Steps (i)-(iii) are repeated until the null hypothesis of cross-sectional independence can no longer be rejected. This sequential testing procedure has unknown statistical size and power properties and may be prone to type II errors.

We therefore choose an alternative way to deal with the issue of multilateral resistance. In terms of the underlying conditional logit model, multilateral resistance, or "the confounding influence that the attractiveness of alternative destinations exerts on the determinants of bilateral migration" (Beine *et al.*, 2016, p. 502), can be interpreted as a violation of the independence of irrelevant alternatives (IIA) property postulating that the relative odds of migrating to two alternative regions s and t depend only on the characteristics of s and t and not on the availability or characteristics of other alternatives. Violations of IIA can arise due to correlations between error terms or correlations between explanatory variables and the error terms (Mokhtarian, 2016).

We therefore check the sensitivity of our results by also using a random parameters logit (RPL) model which relaxes the IIA property.[3] The RPL model can be derived from utility-maximizing behavior by allowing the parameters of a variable z_{isr} to vary over decision makers i, so that $\beta_i \, z_{isr} = (\bar{\mu} + \mu_i) \, z_{isr}$. The parameter β_i is a vector of coefficients for individual i representing *i's* preferences and thus consist of a mean value $\bar{\mu}$ plus an individual-specific deviation from this mean μ_i. The utility function is thus heterogeneous across individuals and the coefficients in β_i are assumed to vary over decision makers according to the density $f(\beta|\theta)$. This so-called 'mixing distribution' describes the distribution of the coefficients β conditional on the parameters θ. Assuming that the main coefficients in β (i.e., the effects of migrant networks, linguistic distance, and their interaction) are normally distributed, the estimated parameters θ are thus the mean and standard deviation of a normal distribution. All other coefficients are modeled as 'fixed' parameters, i.e., parameters whose standard deviation is restricted to zero (Hensher and Greene, 2003).

[3] For an overview see McFadden and Train (2000), Hensher and Greene (2003), and Train (2009).

If the heterogeneity across decision makers is ignored, the individual-specific deviation enters the error term $\varepsilon_{isr} = \mu_i \, z_{isr} + \eta_{isr}$, creating a correlation between observations that share attribute z.[4] With random parameters, the choice probabilities are given by:

$$\Pr(y_{isr} = 1) = \int \frac{\exp\left(V_{isr}\right)}{\sum_{k=1}^{K} \exp\left(V_{isk}\right)} f(\beta) \, d\beta, \tag{2}$$

where $f(\beta)$ is the density function of the parameters β and y_{isr} equals one if individual i from country s chose region r. Estimation of the RPL model is based on the method of maximum simulated likelihood (see Train, 2009, p. 144 for details).

Instead of restoring cross-sectional independence by reducing the variability in the data as in Bertoli and Fernández-Huertas Moraga (2015), our alternative approach thus focuses on modeling the heterogeneity by allowing some of the parameters to vary over decision makers. This requires careful investigations of which variables are modeled as 'random', an issue that we discuss in Section 3.3. Note that while the RPL estimation uses individual-level instead of aggregate data, the approach does not necessarily require individual-level data. Since the coefficients of an individual-level conditional logit model can be estimated using an aggregate-level Poisson model, the reverse must also be true. This in turn implies that an aggregate-level model could, mutatis mutandis, also be estimated using an RPL model to alleviate the issue of multilateral resistance.

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

[4] Cf. Bertoli and Fernández-Huertas Moraga (2015, p. 2), who have already highlighted that "the assumption [...] that the vector of parameters β does not vary across individuals implies that any heterogeneity in the relationship between x_{jk} and U_{ijk} ends up in ϵ_{ijk}, introducing a correlation in the stochastic component of utility across destinations."

data disseminated by Eurostat usually contain only aggregated information on the sending countries, the microdata available to us provides detailed information on migrants' country of birth as well as their region of residence at the NUTS-2 level, which allows the observation of migrant stock on a very small regional level.[5]

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 is modeled for migrants who moved to the EU-15 excluding Ireland (henceforth EU-14) between 1998 and 2007 and who were between 25 and 64 years of age in 2007.[6] Overall, our sample includes 21,315 individual-level observations representing around 7,420,000 recent migrants from 156 sending countries residing in 200 different receiving NUTS-2 regions.[7]

One of our main explanatory variables is the migrant network in region r, which is defined as the stock of migrants from the same sending country s living in region r in 2007 who migrated to country $C(r)$ before 1998:

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

[5] The NUTS classification system (*Nomenclature des unités territoriales statistiques*) is a coherent regional breakdown system administrated by Eurostat. Its purpose is to provide stable regional units over a certain period of time. NUTS-2 regions are based on existing administrative units with an average population size between 800,000 and 3 million. For more information, see http://ec.europa.eu/eurostat/web/nuts/history.

[6] Migration within the EU-15 is not considered because it is governed by a different migration regime than migration to the EU-15 (see Razin and Wahba, 2015). The results, however, are robust to the inclusion of the EU-15 countries. 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.

[7] The total number of observations used in the RPL models is 4,263,000 (= 21,315 individuals × 200 regions) because these models require one observation per alternative for each individual. The number of observations in the PPML models is 31,194 due to the gravity structure (156 sending countries × 200 receiving regions = 31,200 observations) and six sending-receiving combinations with missing information.

Following, amongst others, Ortega and Peri (2009, 2013), Beine *et al.* (2015), and Bertoli and Fernández-Huertas Moraga (2015), we assume a logarithmic form for the network effect and add one to the network size in cases where it is zero to avoid losing observations because of the functional form. The logarithmic specification reflects the assumption of a decreasing marginal utility of migrant networks, so that an increase in the migrant stock has a smaller effect on the probability of choosing a specific region as the size of the network increases.

Unfortunately, the EU-LFS does not allow a more detailed differentiation by year of arrival in the destination country. Despite this shortcoming of the data, three arguments justify our definition of the networks variable: First, it takes some time for networks to be effective; only after previous 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, 2017).

Our second main variable of interest is the linguistic distance between the languages mainly spoken in the home country and the receiving region. In this context, linguistic distance serves as a measure of the proximity of the two languages, thereby indicating the difficulties migrants face in acquiring skills in the host region language (Isphording and Otten, 2014). 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.[8] 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). LD_{sr} is thus defined as the average phonetic similarity between the most commonly spoken language in the sending country and the most commonly spoken language in the receiving region.[9] The interaction between the size of the ethnic network and the linguistic distance, $\text{Network}_{sr} \times \text{LD}_{sr}$, then serves as our variable of main interest. The estimated parameter provides information on the degree of substitutability between the network size and the dissimilarity of the home and host region language.

As control variables, we add further dyad-specific attributes to our estimation model. This includes the geographic distance between the capital of the sending country and the largest city in the receiving region. To capture a possibly diminishing effect for larger distances, this variable enters the estimation in a logarithmic form. As colonial ties can affect the location choice of migrants, we also control for whether the sending and the receiving country share or have ever shared a colonial relationship (Mayer and Zignago, 2011). We further include a dummy variable for whether there is a common official language that is spoken by at least 9 percent of the population in both the sending and receiving countries (Melitz and Toubal, 2014). In addition, we include sending-country fixed effects to control for origin-specific push factors (Ortega and Peri, 2013) and receiving-region fixed effects at the NUTS-2 level to control for destination-specific pull factors. Hence, the representative utility V_{isr} in Eq. (1) is a linear function of sending-country and receiving region specific (dummy) variables, country-pair specific variables (common language and colonial ties), as well as sending country-receiving region specific variables

[8] This measure was first applied to economics by Isphording and Otten (2013) and Isphording and Otten (2014), who analyze the effect of linguistic distance on the language acquisition of immigrants in Germany, Spain, and the US.

[9] 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.

(migrant networks, linguistic distance, and geographic distance), which are assumed to determine the location choice of migrants.[10]

3 Results

3.1 Baseline 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, 2017), 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 the sign and the

Table 1: PPML Estimation of Migration Flows to the EU

	I	II	III	IV	V
	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE	Coef/StdE
Network	0.4033†	0.2758†	0.1300†	0.1151†	0.1114†
	(0.0237)	(0.0469)	(0.0295)	(0.0269)	(0.0265)
LD	−0.0238†	−0.0348†	−0.0334†	−0.0231†	−0.0191†
	(0.0033)	(0.0045)	(0.0029)	(0.0032)	(0.0033)
Network × LD	–	0.0016†	0.0019†	0.0015†	0.0014†
		(0.0005)	(0.0003)	(0.0003)	(0.0003)
ln(distance)	–	–	–	−0.3823***	−0.4592†
				(0.1263)	(0.1258)
Colony	–	–	–	0.9714†	0.4476†
				(0.1138)	(0.1112)
Common off. language	–	–	–	–	0.9558†
					(0.1267)
Constant	−4.0581†	−3.3798***	−5.3086†	−3.4647***	−3.8880***
	(1.0487)	(1.0542)	(1.0986)	(1.3054)	(1.3140)
Sending-country FE	yes	yes	yes	yes	yes
Receiving-region FE	no	no	yes	yes	yes
R^2	0.417	0.409	0.671	0.703	0.708
Observations	31,194	31,194	31,194	31,194	31,194

Notes: – † $p < 0.001$*;* *** $p < 0.01$*;* ** $p < 0.05$*;* * $p < 0.1$. *– Robust standard errors in parentheses. – PPML: Poisson pseudo-maximum-likelihood. – LD: linguistic distance. – Network is defined as the logarithm of the stock of migrants from sending country s living in region r, i.e.,* $Network_{sr} = ln(stock_{sr}^{\leq 1998} + 1)$.

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

significance of the single components of *Network* and *LD* remain stable. This supports the hypotheses that networks and language skills are substitutes in migrants' location choice: networks are more important the larger the linguistic distance between the home country and host region 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 receiving-region 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 sending and the receiving country (Column V).

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.[11] If *LD* equals zero, i.e., if the sending country and the receiving region share a common

Table 2: Change in Odds 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. – LD: linguistic distance. – Network is defined as the logarithm of the stock of migrants from sending country s living in region r, i.e., $Network_{sr} = ln(stock_{sr}^{<1998} + 1)$.

[11] The results are based on our preferred specification shown in Column V of Table 1.

language, a one standard deviation increase in *Network* increases the probability of migrating to that region by about 19 percent. At the 25^{th}-percentile of the distribution of *LD*, however, a similar change in *Network* increases the odds of migrating by about 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 to the region. Similarly, the negative effect of *LD* varies over the distribution of *Network*: While a one standard deviation increase in *LD* is associated with a 22 percent decrease in the probability of migrating to the region at 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*.

The relationship between the other control variables and migrants' location choice is in line with previous literature (see Table 1). The geographic distance between the sending country and the receiving 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).

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 sending and the receiving country as an additional control variable.[12] 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

[12] 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 other covariates remain stable in both size and significance when genetic distance is controlled for. Hence, we can rule out that unobserved cultural differences between the sending and the receiving country are driving our results.

Table 3: PPML Estimation of Migration Flows: Robustness Checks

	I Coef/StdE	II Coef/StdE	III Coef/StdE	IV Coef/StdE
Network	$0.1109^{\dagger}$	-0.1698	–	–
	(0.0265)	(0.1033)		
LD	$-0.0192^{\dagger}$	$-0.0656^{\dagger}$	$-0.0143^{\dagger}$	$-0.0728^{\dagger}$
	(0.0033)	(0.0071)	(0.0024)	(0.0099)
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)	
Linguistic network	–	–	–	-0.1579
				(0.1220)
Linguistic network × LD	–	–	–	$0.0043^{\dagger}$
				(0.0012)
ln(distance)	$-0.4587^{\dagger}$	$-0.4658^{\dagger}$	$-0.8535^{\dagger}$	$-0.4300^{\dagger}$
	(0.1270)	(0.1293)	(0.1385)	(0.1281)
Colony	$0.4438^{\dagger}$	$0.4631^{\dagger}$	$0.5711^{\dagger}$	$0.5061^{\dagger}$
	(0.1113)	(0.1105)	(0.1075)	(0.1111)
Common off. language	$0.9609^{\dagger}$	$1.0273^{\dagger}$	$1.1553^{\dagger}$	$1.0924^{\dagger}$
	(0.1268)	(0.1362)	(0.1308)	(0.1288)
Genetic distance	-0.0210	–	–	–
	(0.2940)			
Constant	-1.5408	0.0888	-1.9284	-0.8117
	(1.2467)	(1.3551)	(1.3655)	(1.4916)
Sending-country FE	yes	yes	yes	yes
Receiving-region FE	yes	yes	yes	yes
Sample LD = 0 incl.	yes	no	yes	no
R^2	0.708	0.675	0.649	0.670
Observations	30,794	30,451	31,194	30,451

Notes: – † $p < 0.001$; *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. – Robust standard errors in parentheses. – PPML: Poisson pseudo-maximum-likelihood. – LD: linguistic distance. – Information on genetic distance is not available for Andorra an the State of Palestine, reducing the sample by 400 observations. – Network is defined as the logarithm of the stock of migrants from sending country s living in region r, i.e., $Network_{sr} = ln(stock_{sr}^{\leq 1998} + 1)$. – Relative network is defined as the stock of migrants from sending country s living in region r divided by the total number of migrants from that sending country living in the EU, i.e., Relative $network_{sr} = (stock_{sr}^{\leq 1998}/stockEU_{s}^{\leq 1998}) \times 100$. – Linguistic network is defined as the logarithm of the stock of migrants living in region r that were born in a country that has the same most common language as the migrant's country of birth s.

Second, we restrict our sample to observations with positive values of *LD*, i.e., we eliminate migration flows between sending countries and receiving regions that have the same 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 II of Table 3. When excluding observations with $LD = 0$, the coefficient of the single component of the network effect turns negative and insignificant. 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 networks 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 can be inferred from Figure 1, which shows the estimated elasticity between networks and migration 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.

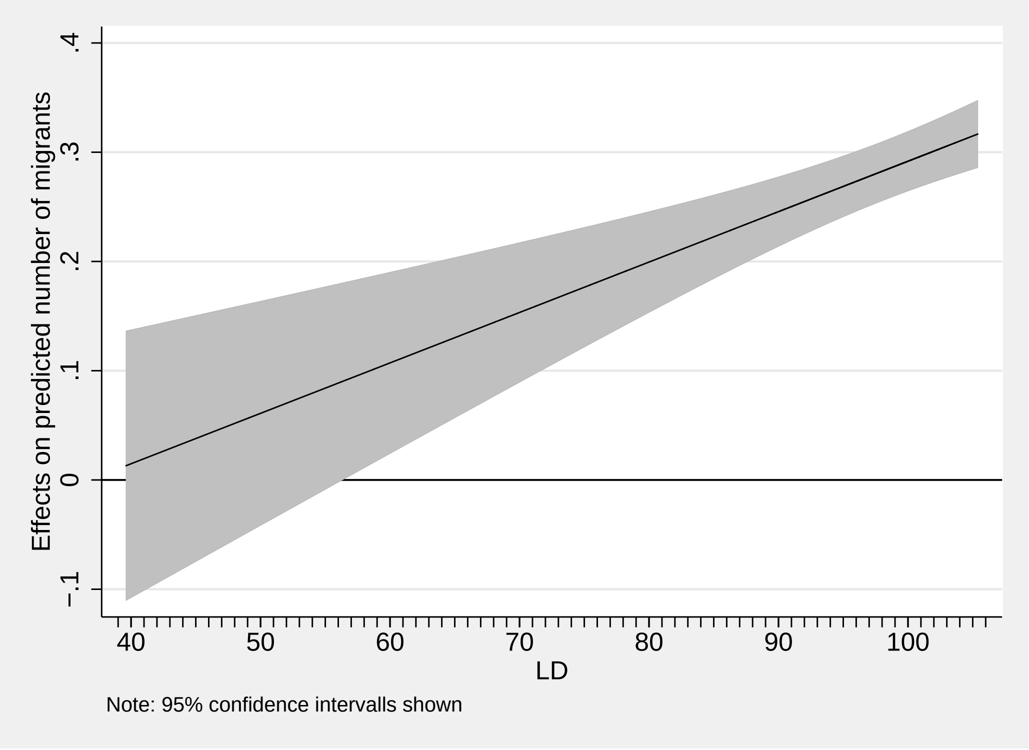

Figure 1: Effect of Network over LD > 0

Notes: – The figure is based on the results shown in Column II of Table 3. – Effect of Network is shown in elasticities. – LD > 0 is effectively bound between 39.6 and 105.4 (see Table A2). – LD: linguistic distance. – Network is defined as the logarithm of the stock of migrants from sending country s living in region r, i.e., $Network_{sr} = ln(stock_{sr}^{\leq 1998} + 1)$.

Lastly, we employ two 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 sending country s living in region r for at least 10 years divided by the total number of migrants from that sending country living in the EU for at least 10 years. The respective estimation results are shown in Column III 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.

Our second alternative measure of the network is based on the idea that it could rather be linguistic networks than same-country networks that matter for migrants' location choice. For a migrant from Ecuador, for example, the number of Spanish-speaking migrants in a region should be an important driver of their location choice, irrespective of whether these migrants are from Ecuador or from any other Spanish-speaking country. To test this idea, we define *Linguistic network* as the number of migrants that speaks the same source-country language, i.e., as the stock of migrants living in region r for at least 10 years that were born in a country that has the same language as the migrant's country of birth s. As the size of the linguistic network should only matter for migrants whose source-country language differs from the language spoken at the destination region, we restrict our sample to observations with $LD > 0$.[13] As is evident from Column IV of Table 3, the interaction effect between linguistic distance and the linguistic network is positive and highly significant, while the single component of the network effect is negative and insignificant. In terms of magnitude, the size of these effects is similar to the respective

[13] Considering the whole sample, we also find a positive, though smaller and hardly significant (10-percent level) interaction effect between linguistic distance and linguistic networks. The estimation results are available from the authors upon request.

effects when using country of birth to define the size of the network (see Column II). This suggests that both ethnic and linguistic networks substitute for linguistic proximity in migrants' location choice.

3.3 Results Based on the Random Parameters Logit Model

As an alternative way to deal with the issue of multilateral resistance to migration, we estimate the underlying choice model using an RPL framework, relaxing the assumption that the vector of parameters does not vary across individuals. We first consider variation in the country preferences of migrants as a source of individual heterogeneity. This approach assumes that the utility from migration to a specific receiving country differs across sending countries due to unobserved dyadic factors unrelated to the already included factors. To model this preference heterogeneity, we include dummy variables for the individual receiving countries whose effects are allowed to vary over decision makers. The RPL model estimates both the mean and the standard deviation of a random parameter. We assume that the deviations from the mean country effects are i.i.d. normally distributed with mean zero and standard deviation σ_β.[14] The random parameters interpretation is formally equivalent to an error components interpretation (Train, 2009, p. 139) where the receiving country dummies represent error terms that create correlations among the utilities of regions within the same country. In this interpretation, the resulting model is analogous to a nested logit model where the regions are nested within countries (Train, 2009, p. 139) or to the model proposed by Bertoli and Fernández-Huertas Moraga (2015) with the receiving countries defined as nests.

Table 4 shows that while almost all of the estimated random parameter means for the country dummies are highly significant relative to the base category (Austria), none of the estimated standard deviations are significant. Since they are also jointly insignificant ($\chi^2(13) = 5.848$, $p = 0.952$), the null hypothesis of no preference heterogeneity for receiving countries across individuals cannot be rejected. The estimated model thus reveals no

[14] In addition, regional fixed effects are included to capture within-country differences in the attractiveness of the regions, which can be interpreted relative to the mean country effect.

Table 4: RPL Estimation of Migration Flows, Heterogeneous Receiving-Country Effects

	Coef/StdE	StdDev/StdE	% $\beta > 0$
Network	$0.1115^{\dagger}$	–	–
	(0.0133)		
LD	$-0.0191^{\dagger}$	–	–
	(0.0015)		
Network × LD	$0.0014^{\dagger}$	–	–
	(0.0002)		
ln(distance)	$-0.4593^{\dagger}$	–	–
	(0.0477)		
Colony	$0.4479^{\dagger}$	–	–
	(0.0596)		
Common off. language	$0.9565^{\dagger}$	–	–
	(0.0711)		
Receiving country dummies (Ref.: AT)			
Country: BE	$2.2499^{\dagger}$	0.0044	100.000
	(0.1465)	(0.0527)	
Country: DE	$2.0016^{\dagger}$	0.0015	100.000
	(0.2569)	(0.0246)	
Country: DK	$2.5675^{\dagger}$	0.0422	100.000
	(0.1433)	(0.0823)	
Country: ES	$2.1735^{\dagger}$	0.0060	100.000
	(0.2064)	(0.0161)	
Country: FI	−0.0071	0.0092	22.148
	(0.3710)	(0.2598)	
Country: FR	$3.1014^{\dagger}$	0.0674	100.000
	(0.1635)	(0.0884)	
Country: GR	0.3123	0.0402	100.000
	(0.1916)	(0.0387)	
Country: IT	$3.0092^{\dagger}$	0.0394	100.000
	(0.1499)	(0.0995)	
Country: LU	$1.0380^{\dagger}$	0.3305	99.916
	(0.1873)	(0.3304)	
Country: NL	0.4041	0.0042	100.000
	(0.2546)	(0.0597)	
Country: PT	$1.0386^{\dagger}$	0.0571	100.000
	(0.2238)	(0.0502)	
Country: SE	$1.9950^{\dagger}$	0.0458	100.000
	(0.1483)	(0.0931)	
Country: UK	0.8834^{***}	0.0442	100.000
	(0.2867)	(0.0479)	
Sending-country FE	yes		
Receiving-region FE	yes		
Observations	21,315 × 200		
Log-likelihood	–28,578.586		

Notes: – † $p < 0.001$; *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. – *Robust standard errors in parentheses. – RPL: Random parameters logit. – LD: linguistic distance. – Network is defined as the logarithm of the stock of migrants from sending country s living in region r, i.e.,* $Network_{sr} = ln(stock_{sr}^{\leq 1998} + 1)$. – *RPL log likelihood simulated using 100 Halton draws.*

evidence in support of a correlation of error terms across regions within the same country. All other coefficients are virtually unchanged compared to the PPML results of Table 1.

Most importantly, the interaction term between the size of the migrant network and linguistic distance does not change in magnitude and significance.

In addition to the estimated random parameter means and standard deviations, Table 4 shows the proportion of the parameters' normal PDF that is above zero. This gives the percentage of the sample for which the parameter is positive. If part of a coefficient's distribution is below zero, the variable constitutes an attractor for some, and a repellent for other individuals. With only two exceptions, the PDFs of the heterogeneous country effects are completely positive. Given that the standard deviations are insignificant, we conclude that most of the country effects are uniformly positive relative to the base category.

In a second step, we estimate an RPL model where we assume that there is heterogeneity across individuals in the preferences for the included dyadic characteristics. In this model, all explanatory variables, including the network, linguistic distance, and the interaction between the two variables, are modeled as normally distributed random parameters. There are good reasons to assume preference heterogeneity for most of the explanatory variables: A common official language may, for example, not be spoken by everyone in the sending country, so that the effect of the variable may not be homogeneous. Likewise, the impact of linguistic distance may depend on (unobserved) individual language skills and ability. Geographical distance may also have a heterogeneous effect due to individual differences in the ability to cover migration costs. Finally, the attractiveness of migrant networks may differ for migrants from ethnically heterogeneous sending countries[15] or if large migrant networks result in statistical discrimination in the receiving region. As above, the model also includes regional fixed effects to control for unobserved region-specific factors that might affect the location decision.

According to the RPL estimation results in Table 5, migrants' preferences for networks and common official language are heterogeneous: the estimated standard deviations of both random parameters are significantly different from zero. As shown in the third column, however, the majority of the two parameter distributions is above zero, indicating

[15] Imagine a sending country comprised of two different ethnic groups where a civil war has driven many members of one ethnic group abroad. Members of this ethnic group will find the networks abroad attractive, while members of the other ethnic group will not.

Table 5: RPL Estimation of Migration Flows, Heterogeneous Dyad-Specific Effects

	Coef/StdE	StdDev/StdE	% $\beta > 0$
Network	$0.2645^{\dagger}$	$0.3533^{\dagger}$	77.300
	(0.0225)	(0.0214)	
LD	$-0.0223^{\dagger}$	0.0018	0.000
	(0.0021)	(0.0038)	
Network × LD	$0.0015^{\dagger}$	$0.0014^{\dagger}$	85.159
	(0.0002)	(0.0004)	
ln(distance)	$-0.4315^{\dagger}$	0.0367	0.000
	(0.0510)	(0.0354)	
Colony	$0.3183^{\dagger}$	0.0468	100.000
	(0.0697)	(0.1293)	
Common off. language	$0.8806^{\dagger}$	0.7938^{***}	86.637
	(0.0833)	(0.2454)	
Sending-country FE	yes		
Receiving-region FE	yes		
Observations	21,315 × 200		
Log-likelihood	–28,339.340		

Notes: – † $p < 0.001$; *** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$. – Robust standard errors in parentheses. – RPL: Random parameters logit. – LD: linguistic distance. – Network is defined as the logarithm of the stock of migrants from sending country s living in region r, i.e., $Network_{sr} = ln(stock_{sr}^{\leq 1998} + 1)$. – RPL log likelihood simulated using 100 Halton draws.

that only few migrants find receiving regions attractive that have small or no networks and/or no common official language. Conversely, we cannot reject the null hypothesis of homogeneous preferences for the linguistic distance, geographic distance, and colonial history variables.[16] In line with the results of the other regressions, we find a statistically significant and positive interaction term between linguistic distance and the size of the migrant network. Although the significant standard deviation of the random parameter implies that the effect of the interaction term is heterogeneous, for a parameter distributed as $N(0.0015, 0.0014^2)$, 85.2 percent of the area under the density function are above zero. This implies that the interaction term is mostly positive, as expected, supporting the result from our baseline model. This kind of heterogeneity could not be identified by using, for example, the PPML or conditional logit model, which supports the decision to estimate the model using RPL.

[16] To check whether the insignificant standard deviation of *LD* is due to *Common official language* being a random parameter, we also estimated a model where *Common official language* is modeled as a fixed parameter. However, the standard deviation of the random *LD* parameter was again insignificant. Results are available from the authors upon request.

4 Conclusion

In this paper, we investigate the role of migrant networks and linguistic proximity in the regional location choice of migrants to the EU. In particular, we analyze whether migrant 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 Labour Force Survey (EU-LFS), which allows us to identify migrants at the regional (NUTS-2) level. Combining this unique dataset with a linguistic distance matrix for a comprehensive set of sending country-receiving region dyads enables us to capture within-country variation in linguistic distance and networks, respectively. This allows us to analyze the regional location choice of migrants at a very disaggregated level, an aspect that has mostly been neglected by the existing literature.

Deriving from the behavioral model, we aggregate the individual data at the bilateral level and estimate the location choices using a Poisson pseudo-maximum likelihood estimator (PPML). To check whether our results are potentially exposed to an omitted variable bias due to multilateral resistance to migration, we also employ a random parameters (mixed) logit (RPL) framework on the individual data. The RPL model relaxes the IIA property and allows for heterogeneous utility functions and can thus be considered as an alternative way to deal with the issue of multilateral resistance to migration.

Our results reveal that both migrant networks and linguistic distance are important determinants of the regional location choice of migrants to the EU. Consistent with the literature on international migration, we find a strong positive network effect and a significant negative effect of linguistic distance. However, we are also the first to show that migrant networks and linguistic proximity represent substitutes in migrants' location choice: migrant networks become more important the larger the linguistic distance between the home country and the host region, and the negative effect of linguistic distance decreases with increasing network size. Altogether, these results suggest that higher migration costs, due to higher language acquisition costs or a smaller network, can be offset by a larger network or by a lower linguistic distance. These results are extremely robust to a number of sensitivity analyses and extensions. Especially, by using the RPL framework to model migrants' location choices, we can show that our results are not biased by multilateral resistance to migration.

Although we do not claim to identify the true 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 provide important insights on international migration flows. Not only do they improve our understanding of the regional location choice of migrants, but also do they have important implications for the possible direction of future migration flows. Over the past years, the EU experienced an unprecedented influx of refugees, most of them fleeing from war and terror in Syria and other countries. If the situation in their home countries does not change substantially in the short-term, many refugees are going to settle permanently in their new destinations. As the settlement of refugees is often influenced by policy decisions out of the control of the refugees, their settlement patterns differ from those of economic migrants, thus creating new ethnic networks. Our findings suggest that such newly established networks will substantially reduce the adverse effects of linguistic barriers for new migrants, thereby increasing the propensity to migrate and settle in locations that are, from a linguistically perspective, very different to the migrant's home country and this way shape future migration flows.

References

Åslund, O. (2005). Now and forever? Initial and subsequent location choices of immigrants. *Regional Science and Urban Economics*, **35** (2), 141–165.

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.

Bauer, T., Epstein, G. S. and Gang, I. N. (2005). Enclaves, language, and the location choice of migrants. *Journal of Population Economics*, **18** (4), 649–662.

Beine, M., Bertoli, S. and Fernández-Huertas Moraga, J. (2016). A practitioners' guide to gravity models of international migration. *The World Economy*, **39** (4), 496–512.

—, 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.

Bleakley, H. and Chin, A. (2004). Language Skills and Earnings: Evidence from Childhood Immigrants. *The Review of Economics and Statistics*, **86** (2), 481–496.

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. (1995). The Endogeneity between Language and Earnings: International Analyses. *Journal of Labor Economics*, **13** (2), 246–288.

— and — (2005). Do Enclaves Matter in Immigrant Adjustment? *City & Community*, **4** (1), 5–35.

— and — (2010). Occupational Language Requirements and the Value of English in the US Labor Market. *Journal of Population Economics*, **23** (1), 353–372.

— and — (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.

Clarke, A. and Isphording, I. E. (2017). Language Barriers and Immigrant Health. *Health Economics*, **26** (6), 765–778.

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

Dustmann, C. and Van Soest, A. (2002). Language and the Earnings of Immigrants. *Industrial and Labor Relations Review*, **55** (3), 473–492.

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.

Hensher, D. A. and Greene, W. H. (2003). The Mixed Logit model: The state of practice. *Transportation*, **30** (2), 133–176.

Isphording, I. E. and Otten, S. (2013). The Costs of Babylon – Linguistic Distance in Applied Economics. *Review of International Economics*, **21** (2), 354–369.

— and — (2014). Linguistic Distance and the Language Fluency of Immigrants. *Journal of Economic Behavior & Organization*, **105**, 30–50.

—, — and Sinning, M. (2014). The Effect of Language Deficiency on Immigrant Labor Market Outcomes in Germany. Mimeo.

Lazear, E. (1999). Culture and Language. *Journal of Political Economy*, **107** (6), 95–126.

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

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

— and Train, K. E. (2000). Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, **15** (5), 447–470.

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

Mokhtarian, P. L. (2016). Presenting the Independence of Irrelevant Alternatives property in a first course on logit modeling. *Journal of Choice Modelling*, **21**, 25–29.

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.

— and — (2017). Network Migration: Do Neighbouring Regions Matter? *Journal of Regional Science*, forthcoming.

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.

Rodríguez-Pose, A. and Ketterer, T. D. (2012). Do local amenities affect the appeal of regions in Europe for migrants? *Journal of Regional Science*, **52** (4), 535–561.

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.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. New York: Cambridge University Press, 2nd edn.

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
Sending-country and receiving-region language pairs			
English Jamaican Creole	39.61	Catalan Swahili (Tanzania)	105.13
Finnish Estonian	47.55	Danish Palestinian Arabic	105.27
Danish Norwegian	47.85	Greek Swazi (Swaziland)	105.39

Notes: – The table shows the five closest and furthest languages toward English and the three closest and furthest sending-country and receiving-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. language	0.093	0.291	0.000	1.000
Genetic distance	0.917	0.725	0.000	2.760
Relative network	0.478	3.147	0.000	100.000
Relative network × LD	38.687	274.228	0.000	$10,098.000$
Linguistic network	2.249	3.451	0.000	12.491
Linguistic network × LD	180.704	293.012	0.000	$1,215.394$
Observations		31,194		

Notes: – LD: linguistic distance. – Information on genetic distance is not available for Andorra an the State of Palestine, reducing the sample by 400 observations. – Network is defined as the logarithm of the stock of migrants from sending country s living in region r, i.e., $Network_{sr} = ln(stock_{sr}^{<1998} + 1)$. *– Relative network is defined as the stock of migrants from sending country s living in region r divided by the total number of migrants from that sending country living in the EU, i.e.,* $Relative\ network_{sr} = (stock_{sr}^{<1998}/stockEU_{s}^{<1998}) \times 100$. *– Linguistic network is defined as the logarithm of the stock of migrants living in region r that were born in a country that has the same most common language as the migrant's country of birth s.*