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

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

[1] *RWI, Rheinisch-Westfälisches Institut für Wirtschaftsforschung*
[2] *University of Salzburg*
[3] *Austrian Institute of Economic Research WIFO*
[4] *Ruhr University Bochum*
[5] *University of California Los Angeles*

This version: August 2015

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 participants at the ERSA congress 2015 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.

Extended Abstract

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), 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 non-French and English speaking immigrants to Canada have a higher probability of settling in areas with a large concentration of the same ethnic group.

This paper therefore analyzes the interaction between migrant networks and linguistic distance in the location decisions of migrants to the European Union. 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:

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

where X_{isr} is a set of control variables specific to migrant i, source country s, region r, and the dyad sr, respectively, and ε_{isr} is a random error term. 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 ε_{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) and Schmidheiny and Brülhart (2011).

The empirical analysis is based on individual level data from a special evaluation of the 2007 European Labour Force Survey (EU-LFS), which provides not only information on the region of residence (at the NUTS-2 level), but also information on the country of birth. 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 between 1998 and 2007.[1] Overall, our data cover about 9 million migrants from 189 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:

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

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). 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. 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 geographic distance (in 1,000 km) between the capital of the source country and the largest city in the host region and control for whether the source and the host country share or have shared a colonial relationship after 1945 (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.

Our main estimation results are shown in Table 1. As expected, we find that the size of the ethnic network has a positive effect on migrants' location choice, while the effect of linguistic distance is negative through all specifications. Regarding our variable of main interest, the interaction between the two variables, we find a positive relationship between the interaction term and migrants' location decision. This supports the hypotheses that networks are more important the larger the linguistic distance between the home and host countries or, stated differently, that the negative effect of linguistic distance is smaller the larger the network size. The positive interaction effect remains after controlling for

[1] Due to missing information on country of birth in the LFS data for Ireland, the set of host countries includes only 14 of the 15 EU member states as of 1998.

[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.

different sorts of fixed effects and for the existence of a common official language between the source country and the host region.

While this result provides a first indication of a substitutive relationship between ethnic networks and linguistic distance, the estimated effect might be liable to estimation bias. First, specification (1) might be inconsistent with the assumptions on the error term ε_{isr}. If X_{isr} 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 (Bertoli and Fernández-Huertas Moraga, 2013, 2015). To address this problem, we will follow Bertoli and Fernández-Huertas Moraga (2015) and 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. In particular, we will use information on language trees to define the nests as a set of regions that belong to the same linguistic group.

Second, the network variable might be endogenous itself. It could be the case that unobservable bilateral components, as for example the cultural proximity between the sourc and the host country, simultaneously affect the stock of migrants, the current flows of new migrants and their selection. To address this problem, we will follow Beine *et al.* (2011) and instrument the size of the ethnic network by the existence of a temporary guest-worker agreement between the source and the host country in the 1960s and 1970s.

References

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

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.

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.

Chiswick, B. R. and Miller, P. W. (2015). *Economics of International Migration 1A*. Oxford and Amsterdam: North Holland.

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

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). The Log of Gravity. *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 Z. P. Frontiers (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. (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.

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.

Tables

Table 1: PPML Estimation of Migration Flows to the EU

Model	I	II	III	IV	V	VI
Network	0.3355**	0.1305**	0.1208**	0.1139**	0.2243**	0.2508**
	(0.0555)	(0.0294)	(0.0267)	(0.0260)	(0.0166)	(0.0189)
LD	−0.0238**	−0.0333**	−0.0277**	−0.0203**	−0.0106**	
	(0.0036)	(0.0029)	(0.0030)	(0.0034)	(0.0025)	
Network×LD	0.0019**	0.0018**	0.0014**	0.0014**		
	(0.0006)	(0.0003)	(0.0003)	(0.0003)		
Common off. lang.				1.0046**	1.0159**	2.0502**
				(0.1397)	(0.1428)	(0.1834)
Network×COL						−0.1023**
						(0.0238)
Control variables	No	No	Yes	Yes	Yes	Yes
Fixed effects	No	Yes	Yes	Yes	Yes	Yes
Dyads	31,394	31,394	31,394	31,394	31,394	31,394
R^2	0.1900	0.6712	0.6861	0.7041	0.6885	0.7098

Notes: – Robust standard errors in parentheses. – ** significant at 1 % level; * significant at 5 % level. – LD: linguistic distance. – COL: common official language. – PPML: Poisson pseudo-maximum-likelihood.