RUHR
ECONOMIC PAPERS

Ingo E. Isphording
Sebastian Otten

The Costs of Babylon – Linguistic Distance in Applied Economics

RWI RUHR UNIVERSITÄT BOCHUM RUB

#337

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, email: W.Leininger@wiso.uni-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Christoph M. Schmidt
RWI, Phone: +49 (0) 201/81 49-227, e-mail: christoph.schmidt@rwi-essen.de

Editorial Office

Joachim Schmidt
RWI, Phone: +49 (0) 201/81 49-292, e-mail: joachim.schmidt@rwi-essen.de

Ruhr Economic Papers #337

Responsible Editor: Thomas K. Bauer

ISSN 1864-4872 (online) – ISBN 978-3-86788-389-4

Ruhr Economic Papers #337

Ingo E. Isphording and Sebastian Otten

The Costs of Babylon – Linguistic Distance in Applied Economics

RWI
RUHR UNIVERSITÄT BOCHUM RUB

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

http://dx.doi.org/10.4419/86788389
ISSN 1864-4872 (online)
ISBN 978-3-86788-389-4

Ingo E. Isphording and Sebastian Otten[1]

The Costs of Babylon – Linguistic Distance in Applied Economics

Abstract

Linguistic distance, i.e. the dissimilarity between languages, is an important factor influencing international economic transactions such as migration or international trade flows by imposing hurdles for second language acquisition and increasing transaction costs. To measure these costs, we suggest to use a new measure of linguistic distance. The Levenshtein distance is an easily computed and transparent approach of including linguistic distance into econometric applications. We show its merits in two different applications. First, the effect of linguistic distance in the language acquisition of immigrants is analyzed using data from the 2000 U.S. Census, the German Socio-Economic Panel, and the National Immigrant Survey of Spain. Across countries, linguistic distance is negatively correlated with reported language skills of immigrants. Second, applying a gravity model to data on international trade flows covering 178 countries and 52 years, it is shown that linguistic distance has a strong negative influence on bilateral trade volumes.

JEL Classification: J24, J61, F22, F16

Keywords: Linguistic distance; immigrants; language; transferability; human capital; international trade

May 2012

1 Ingo E. Isphording, Ruhr-Universität Bochum and RUB Research School; Sebastian Otten, Ruhr-Universität Bochum and RWI. *– The authors are grateful to Thomas K. Bauer, John P. Haisken-DeNew, Ira N. Gang, Julia Bredtmann, Jan Kleibrink, Maren Michaelsen, the participants of the 5th RGS Doctoral Conference, the 2012 Royal Economic Society Annual Conference, and the Society of Labor Economists 2012 Annual Meetings for helpful comments and suggestions. We are also very thankful to Andrew K. Rose for providing much of the trade dataset and Johannes Lohmann for giving us the data of his language barrier index. Financial support from the German-Israeli Foundation for Scientific Research and Development (GIF) is gratefully acknowledged. All remaining errors are our own.* *– All correspondence to: Ingo Isphording, Chair for Economic Policy: Competition Theory and Policy, Ruhr-Universität Bochum, 44780 Bochum, Germany, E-mail: ingo.isphording@rub.de.*

1 Introduction

According to biblical accounts, the Babylonian Confusion once stopped quite effectively the construction of the tower of Babel and scattered the previously monolingual humanity across the world, speaking countless different languages. In economic research, linguistic diversity is believed to be a crucial determinant of real economic outcomes, due to its impact on communication and language skills (see, e.g., Chiswick and Miller 1999), and as accumulated costs affecting international trade flows (see, e.g., Lohmann 2011).

The operationalization of linguistic differences between languages is not straightforward and only few, but problematic, approaches have been undertaken so far. This study proposes to use a measure of linguistic distance developed by linguistic researchers. Linguistic distance is defined as the dissimilarity of languages, including, but not restricted to, vocabulary, grammar, pronunciation, scripture, and phonetic inventories. The *Automatic Similarity Judgement Program* (ASJP) by the German Max Planck Institute for Evolutionary Anthropology offers a descriptive measure of phonetic similarity: the normalized and divided Levenshtein distance. This distance is based on the automatic comparison of the pronunciation of words from different languages having the same meaning. We use this measure in two applications to explain the costs of linguistic differences on the micro- and macro-level. On the micro-level, we use multiple datasets, the 2000 U.S. Census, the National Immigrant Survey of Spain and the German Socio-Economic Panel, to estimate the initial disadvantages due to differences in linguistic origin in the language acquisition of immigrants. On the macro-level, trade-flow gravity models are estimated using the bilateral trade-flow data by Rose (2004) to analyze accumulated costs of linguistic barriers in international trade.

Epstein and Gang (2010) point out that differences in culture, though crucially affecting economic outcomes, are typically treated as a black box in empirical investigations. One main channel of this effect of cultural distance on economic outcomes are differences arising from different linguistic backgrounds. Differences in language are arguably the most visible manifestation of such cultural differences.

Previous studies relied on approaches that measure linguistic distance using average test scores of language students (Chiswick and Miller 1999) or classifications by language families (Guiso, Sapienza and Zingales 2009). Test-score-based approaches assume the difficulty of learning a foreign language for students to be determined by the distance between native and foreign language. Unfortunately, due to data limitations test-score-based measures are only available for the distances towards the English language and are therefore strongly restricted in its use. Approaches using language family trees to derive measures of linguistic distance rely on strong assumptions of cardinality and have to deal with arbitrarily chosen parameters.

Against this background, we contribute to the existing literature in several respects. First, we introduce the normalized and divided Levenshtein distance as an easy and transparently computed, cardinal measure of linguistic distance. We use the general applicability of this measure to broaden the evidence on disadvantages in the language acquisition of immigrants to non-Anglophone countries. Second, we apply the measure in the context of international trade, where language barriers have previously been addressed by controlling for common languages in bilateral trade flows. The Levenshtein distance allows to overcome this very narrow definition of linguistic barriers.

Our results confirm the existence of significant costs of language barriers on the micro- and macro-level. Immigrants coming from a more distant linguistic origin face significantly higher costs of language acquisition. A higher linguistic distance strongly decreases the probability of reporting good language skills. To illustrate the results for immigrants into the U.S., a Vietnamese immigrant coming from a very distant linguistic origin faces an initial disadvantage compared to a German immigrant from a close linguistic origin which is worth of 6 additional years of residence. In the case of accumulated costs on bilateral trade flows, our results indicate that not only a shared common language but also a related but not identical language accelerates trade by lowering transaction costs.

The paper is organized as follows. Section 2 provides a short overview of previous attempts to measure linguistic distance and introduces the Levenshtein distance, discussing its advantages and potential shortcomings. We present our results concerning the explanation of immigrants' language skills in Section 3. The second application, the explanation of international trade flows, is discussed in Section 4. Section 5 summarizes the results and concludes.

2 Measuring Linguistic Distance

2.1 Previous Literature

Linguistic distance is the dissimilarity of languages in a multitude of dimensions, such as vocabulary, grammar, pronunciation, scripture, and phonetic inventories. This multi-dimensionality of linguistic distance makes it difficult to find an appropriate empirical operationalization to be used in applied economic studies.

A very straightforward approach is the evaluation of linguistic distances between languages by counting shared branches in language-family-trees (see, e.g., Guiso, Sapienza and Zingales 2009). This language-tree approach has to deal with strong cardinality assumptions and arbitrarily chosen parameters. Additionally, the approach offers only low variability between different language pairs and is difficult to implement for isolated languages such as Korean.

A widely used approach to measure linguistic distance has been introduced by Chiswick and Miller (1999), who use data on the average test score of U.S. language students after a given time of instruction in a certain foreign language. They assume that the lower the average score, the higher is the linguistic distance between English and the foreign language. Similar measures have been used to analyze the effect of language barriers on international trade (Hutchinson 2005, Ku and Zussman 2010). This test-score-based measurement of linguistic distance relies on strong assumptions. It has to be assumed that the difficulty of U.S. citizens to learn a particular foreign language is symmetric to the difficulty of foreigners to learn English. Further, it has to be assumed that the average test score is not influenced by other language-specific sources. Dörnyei and Schmidt (2001) give an overview of the potential role of intrinsic and extrinsic motivation in learning a second language. Intrinsic motivation, the inherent pleasure of learning a language, and extrinsic motivation, the utility derived from being able to communicate in the foreign language, are likely to differ across languages, but are not distinguishable from the actual linguistic distance in the test-score-based approach.

2.2 The Levenshtein Distance

Drawing from linguistic research, it is possible to derive an operationalization of linguistic distance without strong identification assumptions that underlie previous approaches. The so-called *Automatic Similarity Judgement Program* (ASJP) developed by the German Max Planck Institute for Evolutionary Anthropology aims at automatically evaluating the phonetic similarity between all of the world's languages. The basic idea is to compare 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).

This distance can be interpreted as an approximation of the number of cognates between languages. The linguistic term cognates denotes common ancestries of words. A higher number of cognates indicates closer common ancestries. Although restricting its computation on differences in pronunciation, a lower Levenshtein distance therefore also indicates a higher probability of sharing other language characteristics such as grammar (see Serva 2011). The language acquisition of second language learners is crucially affected by such differences in pronunciation and phonetic inventories, as these determine the difficulty in discriminating between different words and sounds. For a recent overview of the linguistic literature on language background and language acquisition see Llach (2010).

The algorithm calculating the distance between words relies on a specific phonetic alphabet, the ASJPcode. The ASJPcode uses the characters within the standard ASCII[1]

[1] American Standard Code for Information Interchange, keyboard-character-encoding scheme.

alphabet to represent common sounds of human communication. The ASJPcode consists of 41 different symbols representing 7 vowels and 34 consonants. Words are then analyzed as to how many sounds have to be substituted, added, or removed to transfer the one word in one language into the same word in a different language (Holman et al. 2011). The words used in this approach are taken from the so-called 40-item Swadesh list, a list including 40 words that are common in nearly all the world's languages, including parts of the human body or expressions for common things of the environment. The Swadesh list is deductively derived by Swadesh (1952), its items are believed to be universally and culture independently included in all world's languages.[2]

The ASJP program judges each word pair across languages according their similarity in pronunciation. For example, to transfer the phonetic transcription of the English word *you*, transcribed as *yu*, into the transcription of the respective German word *du*, one simply has to substitute the first consonant. But to transfer *maunt3n*, which is the transcription of *mountain*, into *bErk*, which is the transcription of the German *Berg*, one has to remove or substitute each 7 consonants and vowels, respectively.

The following formalization of the computation follows Petroni and Serva (2010). To normalize the distance according to the word length, the resulting number of changes is divided by the word length of the longer word. Denoting this normalized distance between item i of language α and β as $D_i(\alpha, \beta)$, the calculation of the normalized linguistic distance (LDN) is computed as the average across all $i = 1, ..., M$ distances between synonyms of the same item:

$$LDN(\alpha, \beta) = \frac{1}{M} \sum_i D(\alpha_i, \beta_i). \tag{1}$$

To additionally account for potential similarities in phonetic inventories which might lead to a similarity by chance, a global distance between languages is defined as the average Levenshtein distance of words with different meanings:

$$\Gamma(\alpha, \beta) = \frac{1}{M(M-1)} \sum_{i \neq j} D(\alpha_i, \beta_j). \tag{2}$$

The final measure of linguistic distance is then the normalized and divided Levenshtein distance (LDND), which is defined as:

$$LDND(\alpha, \beta) = \frac{LDN(\alpha, \beta)}{\Gamma(\alpha, \beta)}. \tag{3}$$

The resulting measure expresses a percentage measure of similarity between languages, although, by construction, it might take on values higher than 100% in cases in which languages do not even possess these similarities which are expected to exist by chance.

[2]Table A1 in the Appendix shows the list of the 40 words.

Table 1 lists the closest and furthest languages towards English, German, and Spanish. The measurement via the normalized and divided Levenshtein distance is in line with an intuitive guessing about language dissimilarities. Although there is clearly a strong positive correlation between the Levenshtein distance and the test-score-based approach by Chiswick and Miller (1999), the Levenshtein distance offers a higher variability in its measurement and we believe it to be more exact.[3] Some languages are found to be distant according to the Levenshtein distance, but have a comparably low distance using the test-score-based measure, indicating that the test-score-based measure might also entail incentives to learn a foreign language instead of solely measuring linguistic distance.

3 Language Fluency of Immigrants

Language skills of immigrants are known to be a crucial determinant of the economic success of immigrants in the host country labor market. The economic literature concerning the determinants of language fluency of immigrants was initiated by the influential work by Chiswick (1991). Based on this seminal paper, Chiswick and Miller (1995) developed a theoretical human capital framework of host country language skill acquisition. In this framework, linguistic distance is a crucial determinant of language skills by lowering the efficiency of learning a language and inducing higher learning costs. This theoretical implication has been subsequently tested for various countries using the test-scores-based measure by Chiswick and Miller (1999). Due to its exclusive availability to the English language, these applications have been restricted to studies concerning the immigration to English-speaking countries such as the U.S. or Canada (Chiswick and Miller 2005). This restriction does not hold for the Levenshtein distance as a measure of linguistic distance, which is not restricted to any home or host country, and may therefore be applied to a broader range of countries.

This feature allows for providing evidence on the relationship between linguistic distance and language fluency in an international perspective. In doing so, we utilize data from three different sources. First, we use data from the 2000 U.S. Census to apply both the test-score-based measure by Chiswick and Miller (1999) and the Levenshtein distance to the same dataset. To compare the influence of linguistic distance across different countries, we additionally use data from the German Socio-Economic Panel (SOEP), and the National Immigrant Survey of Spain (NISS).

The U.S., Germany, and Spain have very different migration histories that make an international comparison worthwhile. The United States have been an immigrant country since its foundation and currently a legal permanent residence status is granted to about 1 million immigrants per year. In 2000, this immigration flow consisted mainly of immigrants

[3]Figure A1 in the Appendix shows the relationship between both measure.

from other North-American countries (40%, including 21% from Mexico), followed by Asian (32%) and European immigrants (15%) (U.S. Department of Homeland Security 2010). These inflows are also resembled in the stocks of the immigrant population. In the 2000 U.S. Census, 11% of the population of the U.S. were foreign-born.

Neither does Germany have such a long-running immigration history as the U.S., nor can it look back on an extensive colonial history as Spain. Mass immigration started off only shortly after World War II with large waves of ethnic German expellees, followed by the so-called "guestworker"-programs aimed at attracting mainly unskilled workers from Mediterranean countries such as Turkey, Yugoslavia, Italy, or Spain. These two first waves of immigration were followed by a strong immigration phase by family re-unification during the 1970s and 1980s. The third large wave of immigration consists of immigrants and Ethnic Germans from former Soviet states during the 1990s (Bauer et al. 2005). Compared to the U.S. and Spain, Germany has a very old immigrant population with long individual migration histories. In 2009, 10.6 million (approx. 13%) of the German population have immigrated after 1949, 3.3 million as Ethnic Germans. This third immigration wave was accompanied by large numbers of refugees and asylum seekers from Ex-Yugoslawia. The major part of the immigrant population is born in EU member states (32%), followed by 28% from Turkey and 27% from former members of the Soviet Union.

Although Spain has a long-running colonial history, it is a comparably young immigration country. After large waves of emigration until the 1970s, net immigration began in the early nineties, and accelerated considerably during the last 20 years. Between 1997 and 2007, the number of migrants increased by around 700%, initially including mostly migrants from Africa and Western Europe. Nowadays, the majority of immigrants comes from Latin America and, since the EU enlargement, increasingly from Eastern Europe. Today, about 10% or 4.5 million of the population in Spain are foreign-born (see Fernández and Ortega 2008).

3.1 Data and Method

Our data are restricted to male immigrants who entered the respective country after the age of 16 and are younger than 65 and who do not speak the host country language as their first language. The sample drawn from the 1%-PUMS (Public Use Microdata Series) 2000 U.S. Census file consists of 59,889 individuals. Similar data is extracted from the German Socio-Economic Panel, a long-run longitudinal representative study. Using cross-sectional data from 2001, the sample consists of 675 male immigrants.[4] The National Immigrant Survey of Spain, conducted in 2007, also offers comprehensive cross-sectional information

[4]For further information about the SOEP see Haisken-DeNew and Frick (2005). The SOEP data was extracted by using the Stata-add-on PanelWhiz (Haisken-DeNew and Hahn 2006).

on the socio-economic characteristics and migration history of immigrants.[5] The sample includes 2,513 male immigrants.

All datasets include self-reported assessments of language fluency ranging over four and five categories, respectively, which are recoded into dichotomous measures, where 1 means "Good" or "Very Good" language skills and 0 is associated with all lower values. This variable serves as a the dependent variable in Probit regressions. This dichotomization decreases the probability of misclassification, which would lead to biased estimates in the case of Probit models, as pointed out by Dustmann and van Soest (2001). Moreover, it avoids dealing with violated proportional odds assumptions in the case of Ordered Probit models, as discussed by Isphording and Otten (2011). Further, the recoding enhances the comparability of the estimations between the different datasets and to previous approaches, as e.g. Chiswick and Miller (1999).

Denoting this dichotomized indicator variable of host country language skills as our dependent variable y_i, the estimated probability of reporting good language skills can be specified as:

$$Pr\left[y_i = 1 |\, LD_i,\, X_i\right] = \Phi\left(\beta_0 + \beta_1\, LD_i + X_i\, \gamma\right), \tag{4}$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. LD is the linguistic distance between the native language and the host country language, the parameter β_1 is our main parameter of interest, the disadvantage by linguistic origin in the language acquisition process.

The explanatory variables X_i are chosen to ensure the highest possible comparability between the regressions. The main variable of interest is the measure of linguistic distance introduced in Section 2.2. The normalized and divided Levenshtein distance enters the specifications not as an absolute value but as a percentile measure, indicating the position of each individual in the overall distribution of the linguistic distance. As such, we ensure a certain level of comparability between the different ways of measuring linguistic distance.

As additional control variables, all three datasets offer comparable information on the age at migration, years since migration, years of education, marital status, number of children and an indicator variable denoting a former colonial relationship between home- and host country. We additionally include the distance between capitals in kilometers to proxy migration costs. For the 2000 U.S. Census we include some additional regional information about living in a non-metropolitan area, living in the Southern states and the share of the minority speaking the language of the individual. For Germany, we include a dummy for coming from a neighboring country. We control for refugee status (U.S. and Germany) and political reasons for migration (Spain), respectively. The U.S. data

[5] For further information about the NISS see Reher and Requena (2009).

further includes information about having been abroad 5 years ago, while the German data includes information on having family abroad. This information serves as a proxy for return migration probability. Finally, each specification includes 17 world-region dummies to account for potential cultural differences correlated with linguistic distance.

Sample means of the included variables are reported in Table 2. They show significant differences across the datasets, related to the different migration histories summarized above. Immigrants in Germany display the highest number of years since migration, as the sample consists in large parts of former guestworkers who immigrated during the 1960s and early 1970s. The German immigrant population also has the lowest mean education, but a higher share of married couples and a higher number of children, which might partly be due to the higher average age. The low average distance to the home country indicates the high share of guestworkers and immigrants from Eastern and Southern Europe. In contrast, both immigrants to Spain and the U.S. have a high average distance to the home country, as many immigrants come from overseas. Spain has the youngest immigrant population, resembling its relatively short immigration history starting off in the 1990s. Each dataset has a comparable share of "Good" or "Very Good" host country language skills of around half of the sample.

3.2 Results

Table 3 lists the results of the Probit regressions across datasets, reported as marginal effects evaluated at the mean of the covariates. Columns (1) and (2) show the results for the U.S. data, using the test-score-based measure and the Levenshtein distance, respectively. Column (3) shows the results for the German SOEP data, and column (4) for the Spanish NISS data.

The results confirm a significantly negative effect of linguistic distance on the probability of reporting good or very good language abilities in the host country language throughout all estimations. For the U.S., the effects for the test-score-based measure and the Levenshtein distance are qualitatively comparable. The effect is lower, however, when applying the Levenshtein distance.

To illustrate the effect of linguistic distance, we can look at the additional amount of years of residence that make up for an initial disadvantage by linguistic origin. This amount of years of residence can be calculated by equating the marginal effect of years since migration with the marginal effect of a certain difference in linguistic origins and solving for the years since migration. In the U.S., the initial disadvantage of an immigrant with a distant linguistic origin, e.g. a Vietnamese who is in the 97^{th} percentile of the distribution of linguistic distance, compared to an immigrant with close linguistic origin, e.g. a German who is in the 1^{st} percentile, is worth around 6 years of additional residence. For a Turk (79^{th} percentile), the largest immigrant group in Germany, the disadvantage

compared to a linguistically closer Dutch migrant (3^{rd} percentile) is worth 8 years of residence.

Switching the measure of linguistic distance in the U.S. data from the test-score-based to the normalized and divided Levenshtein distance does not qualitatively affect the coefficients of the control variables. The coefficients are in line with previous studies and theoretical predictions. We see a positive impact of education, at around 5 percentage points for the U.S. and Germany and around 3 percentage points for Spain. The initial negative effect of age at migration decreases over time. Being married and having children is associated with higher language skills. The signs of these relationships are stable across all datasets, with the exception of lower language skills for immigrants with children in Spain. Being born in a former colony has a strong positive effect for both immigrants in the U.S. and in Spain. In Germany those immigrants from a neighboring country report higher language skills. Refugees in the U.S. and in Germany report lower average language skills compared to immigrants without refugee status.

4 International Trade

Costs imposed by linguistic barriers can also be found on the macro-level. The trade-increasing effect of a common language is an undisputed fact in international economics. It is intuitive that trade between countries with a common language is cheaper than between countries with different languages. In their survey article, Anderson and van Wincoop (2004) report an estimate of the tax equivalent of "representative" trade costs for industrialized countries of about 170%. Of these, language-related barriers account for 7 percentage points, which is similar in magnitude to policy barriers and information costs.

The question is whether and to what extent the dissimilarity between two languages matters if trading partners do not share a common language. Certainly, a range of dominating languages (English in the Western countries, Russian in Eastern Europe, French in Africa, and Spanish in Latin America) plays a major role in international trade. Especially the role of English as a lingua franca has been addressed by Ku and Zussman (2010). However, in the development of longer-term business partnerships, the crucial variable of interest is the linguistic knowledge in the trade partner's home country language (Hagen, Foreman-Peck and Davila-Philippon 2006), captured by the direct linguistic distance between the dominant languages of the trade partners.

The method of choice in examining determinants of international bilateral trade is the gravity model first proposed by Tinbergen (1962). The basic theoretical gravity model assumes that the size of bilateral trade between any two countries depends on a function of each country's economic size measured by (log of) GDP. Trade costs in their simplest form are approximated by the distance between the trading countries (Anderson and van

Wincoop 2004). Extensions are proxies for trade frictions, such as the effect of trade agreements (McCallum 1995), and cultural proximity (Felbermayr and Toubal 2010).

To incorporate language-related barriers into these gravity models, common empirical practice is to use an indicator variable that equals 1 if two countries share the same official language and 0 otherwise (see Anderson and van Wincoop 2004). While most studies employ the former approach, Melitz (2008) goes beyond official languages and develops two different measures. The first measure depends on the probability that two randomly chosen individuals from either country share a common language spoken by at least 4% of both populations. The second measure is an indicator variable that equals 1 if two countries have the same official language or the same language is spoken by at least 20% of the populations of both countries.

These measures share the shortcoming that they only look at whether countries share the same language, but do not account for heterogeneity in the degrees of similarity between languages. The degree of similarity, however, is likely to affect trade costs, e.g. by lower costs of learning the trade partner's language or by lowering translation costs (Hagen, Foreman-Peck and Davila-Philippon 2006). Our results of Section 3.2, showing that linguistic barriers crucially affects second language acquisition, lend further support to this hypothesis. Moreover, lower host country language skills diminish the ability of immigrants to promote trade and commerce between their host country and their country of origin (Hutchinson 2005).

The only two approaches we know of that take similarities and differences between a multitude of languages into account are the ones by Hutchinson (2005) and Lohmann (2011). By relying on the measure by Chiswick and Miller (1999), Hutchinson's approach is restricted to distances towards English. Lohmann (2011) uses data from the World Atlas of Language Structures (WALS; see Dryer and Haspelmath 2011) to construct an index of 139 potentially shared linguistic features between languages. Similar to our application, he applies this index to explain international trade flows using data from Rose (2004). This approach counts shared language features within language pairs and builds up a language features index normalized to the interval of $[0; 1]$, where 0 means sharing all features.[6]

4.1 Data and Method

To ensure a high degree of comparability with the previous literature, we use a widely accepted empirical methodology and a standard dataset of bilateral trade flows. The dataset constructed by Rose (2004) has been widely used previously by Melitz (2008), Ku

[6]The measure by Lohmann (2011) is assigned at the country-level using the most widely spoken official language of each country. In the Spanish and U.S. micro-data, we can rely on a more detailed assignment using information on the mother tongue of each individual. This makes it unfeasible to include this alternative measure in the micro-data regressions in Section 3.

and Zussman (2010), and Lohmann (2011).[7] The sample covers bilateral trade between 178 countries over the years 1948 to 1999 leading to 234,597 country-pair-year observations. The variables of interest are Rose's binary common language variable, two versions of linguistic distance between trading partners' languages as measured by the Levenshtein distance, and finally the linguistic features index calculated by Lohmann (2011).

The Levenshtein distance is computed for every country-pair in the dataset. In monolingual countries we assign the respective native language to the country. In multi-lingual countries, the most prevalent native language is assigned, which was identified using a multitude of sources, including CIA's World Factbook, encyclopedias, and Internet resources.[8] To analyze the sensitivity of the results with respect to the measurement of linguistic distance, we calculate an alternative specification of the Levenshtein distance, replacing the most prevalent language with the prevailing lingua franca in a country. A lingua franca is defined as a language typically used to enable communication between individuals not sharing a mother tongue. These languages are often third languages, which are widely spoken in a particular regional area and are not necessarily an official language.[9] Subsequently, we compare the effect of these two definitions of the Levenshtein distance with the approach by Lohmann (2011).

Descriptive statistics of the variables used in the empirical analysis are shown in Table 4. The average Levenshtein distance decreases from 90.3 to 75.1 when we use lingua francas instead of the most prevalent native language to calculate the linguistic distance. This indicates that lingua francas may have come into existence to decrease costs imposed by language barriers in the first place. Following Rose' definition, 22.3% of the country-pairs share a common language. This quite high share relies on a very broad definition of official languages by Rose. For example, even country pairs such as the U.S. and Denmark or France and Egypt are coded to have the same language. Using the Levenshtein distance, only 4.7% of the country-pairs show a distance of zero, which is equivalent to sharing a common language, increasing to 18.4% for the Levenshtein distance measure based on lingua francas. The linguistic features index by Lohmann (2011) is zero for 9.4% of the country-pairs, meaning that both languages share all linguistic features considered.

We use the gravity model to estimate the impact of language barriers on trade between pairs of countries. The model has a long record of success in explaining bilateral trade flows and becomes the standard model for applied trade analysis. Following Rose (2004), we augment the basic gravity equation with a number of additional variables that affect

[7]The data and their sources are explained in detail in Rose (2004) and posted on his website. We additionally account for errors identified by Tomz, Goldstein and Rivers (2007).

[8]For example, we use English as the native language in the United Kingdom, because it is a monolingual country and English is the national language. In a multi-lingual country such as Canada we use English instead of French, because English is the most prevalent native language. A comprehensive index of assigned languages with further explanations is available upon request.

[9]For example, we use Russian as lingua franca for most countries of the former Soviet Union.

trade in order to control for as many determinants of trade flows as possible. Our empirical strategy is to compare trade patterns for trading partners with different language barriers using variation across country-pairs. If a common language or a high similarity between languages has a positive effect on trade, we expect to observe significantly higher trade for these country-pairs than for others. We compare three different specifications. First, we adopt the original specification by Rose (2004) including an indicator variable for country-pairs sharing the same language. This basic approach is then augmented by the Levenshtein distance and the language features index by Lohmann (2011). The exact specification of the gravity model is:

$$\begin{aligned} ln\, X_{ijt} = {} & \beta_1\, ln\, (Y_i Y_j)_t + \beta_2\, ln\, Dist_{ij} + Z_{ijt}\, \kappa + \gamma_1\, LangBar_{ij} \\ & + \textstyle\sum_i \delta_i I_i + \sum_j \theta_j J_j + \sum_t \phi_t T_t + \varepsilon_{ijt}, \end{aligned} \tag{5}$$

where the dependent variable X_{ijt} denotes the average value of real bilateral trade between country i and country j at time t, mainly influenced by the "mass" of both economies, indicated by the product of their GDP denoted by Y, and the distance in log miles. Z is a vector of control variables, including population size, geographic characteristics such as sharing a land border, number of landlocked countries, number of island nations in the country-pair (0, 1, or 2), the area of the country (in square kilometers), and colonial relationships. Further, it is controlled for member and nonmember participation in the GATT/WTO (one or both countries), same currency, regional trade agreements, and being a GSP beneficiary.[10]

The main coefficient of interest is γ_1. It measures the effect of the different language barriers variables ($LangBar$) on international trade. If both countries share a common language, γ_1 should be positive; if instead one of the linguistic distance measures is used, the effect of γ_1 on trade should be negative. A comprehensive set of country and year fixed effects is included in the specification to control for any factor affecting trade that is country (e.g. stock of migrants, foreign language knowledge) or time specific (e.g. common shocks and trends).[11] The gravity model is estimated by ordinary least squares (OLS) with robust standard errors clustered on the country-pair level.

[10] More details are given in Rose (2004), the source for all variables except the linguistic distance measures. In the assignment of GATT/WTO rights and obligation we follow Tomz, Goldstein and Rivers (2007) and impose the restriction that formal membership has the same effect as nonmember participation.

[11] Recent empirical work on the determinants of bilateral trade increasingly relies on panel data techniques that account for country-pair instead of exporter and importer specific fixed effects. Country-pair fixed effects control for the impact of any time-invariant country-pair specific determinant such as bilateral distance or common language. However, this comes at the cost of not being able to estimate the effect of the language barrier variables, our variables of interest, on bilateral trade.

4.2 Results

Table 5 summarizes the results of Eq. (5). For the sake of brevity, the estimated coefficients for the time- and country-fixed effects are omitted from all tables.

In the first column, we reproduce the benchmark specification from Rose (2004) based on his measure of common language augmented with country fixed effects. Rose' model confirms the hypothesis of a significant positive effect of common language on bilateral trade. Sharing a common language is found to raise trade by about $(exp(0.274) - 1 \approx)$ 31.5%. Still, this result might be biased by the very broad definition of having a common language.

The question we want to answer is whether language barriers affect trade above and beyond the simple effect of sharing a common language. Therefore, our ensuing specifications examine how the results change when we employ the linguistic distance measures instead of the common language variable.

The second column shows our preferred model. We replace Rose' common language variable with our default Levenshtein distance measure. We find significantly lower trade when the Levenshtein distance between both countries in a dyad increases. The coefficient indicates that a country-pair trades about $(exp(-0.006) - 1 \approx)$ 0.6% less if the Levenshtein distance increases by one unit. To illustrate the magnitude of this effect, we note that the 75^{th} percentile of the Levenshtein distance in our sample is 99.93 (roughly the distance between English and Japanese) and the 25^{th} percentile is 92.95 (roughly the distance between English and Russian). The estimate in column 2 implies that an increase from the 25^{th} to the 75^{th} percentile in the Levenshtein distance decreases bilateral trade by approximately 4.1%.[12]

In multi-lingual countries, the assignment of languages to countries is difficult. To show that our findings are not a result of a particular assignment of languages to countries, the estimation results with the Levenshtein distance measure based on lingua francas are presented in column (3). The key result that the Levenshtein distance has a statistically and economically significant negative effect on bilateral trade is robust. However, the effect decreases by 50%, maybe due to the lower variability of the alternative Levenshtein distance. Additionally, lingua francas are purposely chosen to lower transaction costs. Therefore, we should expect a smaller effect on trade when taking the lingua francas into account.

Next, column (4) shows the results of Lohmann's linguistic features index as a measure of common language. Due to a restricted data availability, the linguistic features index is

[12]To examine whether the effect of the Levenshtein distance on bilateral trade only builds on the grounds of sharing or not sharing a common language and not on the linguistic distance between different languages, we estimate models 2-4 with subsamples excluding country-pairs with no language barrier in the corresponding measure, i.e. a linguistic distance of zero. Table A3 in the Appendix provides the results. Regarding both versions of the Levenshtein distance measure, they become even lager in magnitude and are stable in significance, while the linguistic features index becomes distinctly smaller in magnitude and significance.

only computable for a subsample of 227,145 country-pairs.[13] The coefficient reveals that a pair of countries trade about $(exp(-0.618) - 1 \approx)$ 4.6% less if the linguistic features index increases by 0.1 units (corresponding to a 10% decrease in common linguistic features).

To compare the influence of the language features index by Lohmann (2011) to the Levenshtein distance, we compute the elasticity and the marginal effect multiplied by the interquartile range of the linguistic features index. Increasing the linguistic features index by 1% decreases bilateral trade by about 0.3% compared to a 0.6% decrease in case of the Levenshtein distance. Moving up the distribution of the linguistic features index from the lower to the upper quartile decreases trade between countries by $(exp(-1.465) - 1 \approx)$ 7.7%. The results show a larger effect for the Levenshtein distance with regard to elasticities. Since the distribution of the Levenshtein distance is right-skewed, the value of the interquartile range is smaller compared to the interquartile range of the linguistic features index. As a result, the effect of the linguistic features index becomes larger than the effect of the Levenshtein distance. In summary, the empirical analysis provides evidence that according to both measures linguistic distance has a statistically and economically significant negative effect on bilateral trade flows.

The estimated coefficients of the control variables confirm the traditional results of gravity trade equations. The indicators for whether one or both countries in the dyad participated in the GATT/WTO have significantly positive coefficients. The respective coefficients are comparable to those reported by Tomz, Goldstein and Rivers (2007). Countries that are farther apart trade less, while countries belonging to the same regional trade association, belonging to the same GSP, or sharing a currency trade more. Islands or landlocked countries trade less, while countries sharing a land border trade more. Economically larger and richer countries trade more, as do physically larger countries. A shared colonial history encourages trade as well. These estimation results are both statistically and economically significant and in line with estimates from previous literature.

As compared to the first specification, the application of the Levenshtein distance measure does not considerably affect the magnitude or significance of the other independent variables. All variables show the expected results. However, the coefficient of common colonizer increases by about 10 percentage points, indicating that the effect of cultural ties is underestimated in the traditional gravity model. During the colonization period, colonizers created new institutions such as the legal and administrative system in their colonies. These institutions impose policies and law enforcement, thereby determining the formal and informal rules in commerce. Since international transactions between countries with different or poorly developed institutional settings involve high transactions costs, colonial ties between countries that had the same colonial history and therefore established

[13] To check for sample selection we additionally estimated models 1-3 restricted to the same subsample. Table A4 in the Appendix shows the results. The estimates regarding the language variables remained stable in magnitude and significance.

a similar institutional system, facilitate bilateral trade flows. Despite the fact that the colonizers' languages became the official languages in the colonies and represent one of the official languages in most former colonies even today, a large part of the population failed to achieve an acceptable degree of knowledge in these languages (see, e.g., Lewis 2009). Hence, using information on common official languages in a country-pair to estimate trade flows, in particular between countries with a common colonizer, might underestimate the effect of colonial ties and overestimate the relation of the common colonizers language on trade pattern. In summary, a common colonizer promotes trade between these countries because of establishing a similar institutional setting; an effect that might be hidden when not controlling properly for linguistic heterogeneity.

5 Conclusion

This study is concerned with the operationalization of linguistic distance between languages and the estimation of arising costs of linguistic barriers on the micro- and macro-level. Linguistic barriers are strong obstacles in the realization of free worldwide factor movements. The operationalization of linguistic barriers in applied economic studies is not straightforward and makes it necessary to rely on interdisciplinary approaches drawing heavily from linguistic research. Our measure for linguistic distance is based on the *Automatic Similarity Judgment Program* (ASJP) by the German Max Planck Institute for Evolutionary Anthropology. The linguistic distance is computed as a function of phonetic similarity of words (a Levenshtein distance) from different languages having the same meaning. It can be used as an approximation of the historical difference in languages and is therefore also correlated to differences in other dimensions of dissimilarity, such as grammar or vocabularies.

Compared to the previous approach by Chiswick and Miller (1999), which measures linguistic distance by using average test-scores of second language students, the Levenshtein distance has some advantages. It is available for any pair of the world's languages (instead of being only applicable for the distance towards English). Additionally, it is not influenced by other extrinsic or intrinsic incentives for learning a foreign language, and should deliver an unbiased approximation of the dissimilarity between languages.

The measurement of linguistic distance is used in two applications, the language acquisition of immigrants and language barriers in bilateral trade flows. Following the widely accepted rational choice framework of language acquisition (see, e.g., Chiswick and Miller 1995, Esser 2006), linguistic distance affects second language skills by lowering the initial efficiency, thereby imposing higher costs of learning a foreign language. Following previous work that shows such a negative relationship for English-speaking countries, we broadened the evidence for other countries by applying the measure in estimations

using U.S., German, and Spanish individual micro-data. The results confirm a strong significantly negative effect of linguistic distance on immigrant language skills. The initial disadvantage due to distant linguistic origin is worth several years of additional residence. As such, the linguistic distance is able to explain a large part of language skill heterogeneity in immigrant populations. The considerable hurdles for language acquisition on the micro-level might explain the lower migration rates between linguistically distant countries, as analyzed by Adsera and Pytlikova (2012).

To additionally look at how these effects on the micro-level accumulate to costs of linguistic barriers on the macro-level, we apply the Levenshtein distance in the setting of international trade. Linguistic proximity is believed to enhance trade flows between countries by lowering costs imposed by language barriers, e.g. translation or information costs. Using a comprehensive dataset of bilateral trade flows by Rose (2004), we estimate a standard gravity model using the Levenshtein distance as an additional explanatory variable and compare this approach to a previous approach based on shared linguistic features by Lohmann (2011). The results provide new and strong evidence indicating that language barriers affect trade above and beyond the simple effect of sharing a common language. Moving up the distribution of the Levenshtein distance from the lower quartile (roughly the distance between English and Russian) to the upper quartile (roughly the distance between English and Japanese) decreases trade between countries by about 4.1%.

Taken together, this study suggests an important role of language differences in economic transactions. The results show the significant economic costs of linguistic heterogeneity on the individual and aggregated level. The Levenshtein distance offers a simple and comprehensive way to control for this heterogeneity in a large range of applications in empirical economics and thereby circumvents potential pitfalls by decreasing the degree of unobserved heterogeneity in the data.

References

Adsera, Alicia, and Mariola Pytlikova. 2012. "The Role of Language in Shaping International Migration." IZA Discussion Papers No. 6333.

Anderson, James E., and Eric van Wincoop. 2004. "Trade Costs." *Journal of Economic Literature*, 42(3): 691–751.

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. "Adding typology to lexicostatistics: A combined approach to language classification." *Linguistic Typology*, 13(1): 169–181.

Bauer, Thomas K., Barbara Dietz, Klaus F. Zimmermann, and Eric Zwintz. 2005. "German Migration: Development, Assimilation, and Labour Market Effects." In *European Migration*. ed. Klaus F. Zimmermann, 197–261. Oxford: Oxford University Press.

Chiswick, Barry R. 1991. "Speaking, Reading, and Earnings among Low-Skilled Immigrants." *Journal of Labor Economics*, 9(2): 149–170.

Chiswick, Barry R., and Paul W. Miller. 1995. "The Endogeneity between Language and Earnings: International Analyses." *Journal of Labor Economics*, 13(2): 246–288.

Chiswick, Barry R., and Paul W. Miller. 1999. "English language fluency among immigrants in the United States." In *Research in Labor Economics*. Vol. 17, ed. Solomon W. Polachek, 151–200. Oxford: JAI Press.

Chiswick, Barry R., and Paul W. Miller. 2005. "Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages." *Journal of Multilingual and Multicultural Development*, 26(1): 1–11.

Dörnyei, Zoltán, and Richard Schmidt. 2001. *Motivation and second language acquisition*. Vol. 23 of *Technical Report/Second Language Teaching & Curriculum Center* Honolulu, HI: Univ. of Hawaii.

Dryer, Matthew S., and Martin Haspelmath. 2011. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library.

Dustmann, Christian, and Arthur van Soest. 2001. "Language Fluency and Earnings: Estimation with Misclassified Language Indicators." *Review of Economics and Statistics*, 83(4): 663–674.

Epstein, Gil S., and Ira N. Gang. 2010. "Migration and Culture." In *Frontiers of Economics and Globalization: Migration and Culture*. Vol. 8, ed. Gil S. Epstein and Ira N. Gang. Emerald Group Publishing Limited.

Esser, Hartmut. 2006. "Migration, Language and Integration: AKI Research Review 4." Social Science Research Center Berlin, Berlin. http://bibliothek.wz-berlin.de/pdf/2006/iv06-akibilanz4b.pdf.

Felbermayr, Gabriel J., and Farid Toubal. 2010. "Cultural proximity and trade." *European Economic Review*, 54(2): 279–293.

Fernández, Cristina, and Carolina Ortega. 2008. "Labor market assimilation of immigrants in Spain: employment at the expense of bad job-matches?" *Spanish Economic Review*, 10(2): 83–107.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2009. "Cultural Biases in Economic Exchange?" *Quarterly Journal of Economics*, 124(3): 1095–1131.

Hagen, Stephen, James Foreman-Peck, and Santiago Davila-Philippon. 2006. "ELAN: Effects on the European Economy of Shortages of Foreign Language Skills in Enterprise." Brussels: European Commission. http://ec.europa.eu/languages/languages-mean-business/files/elan-full-report_en.pdf.

Haisken-DeNew, John P., and Joachim R. Frick. 2005. "Desktop Companion to the German Socio-Economic Panel (SOEP): Version 8.0." German Institute for Economic Research, Berlin. http://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.38951.de/dtc.409713.pdf.

Haisken-DeNew, John P., and Markus Hahn. 2006. "PanelWhiz: A Flexible Modularized Stata Interface for Accessing Large Scale Panel Data Sets." Essen. http://www.panelwhiz.eu/docs/PanelWhiz_Introduction.pdf.

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dirk Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov. 2011. "Automated Dating of the World's Language Families Based on Lexical Similarity." *Current Anthropology*, 52(6): 841–875.

Hutchinson, William K. 2005. ""Linguistic Distance" as a Determinant of Bilateral Trade." *Southern Economic Journal*, 72(1): 1–15.

Isphording, Ingo E., and Sebastian Otten. 2011. "Linguistic Distance and the Language Fluency of Immigrants." Ruhr Economic Papers No. 274.

Ku, Hyejin, and Asaf Zussman. 2010. "Lingua franca: The role of English in international trade." *Journal of Economic Behavior & Organization*, 75(2): 250–260.

Lewis, Paul M. 2009. *Ethnologue: Languages of the World.* 16th ed. Dallas, TX: SIL International.

Llach, Agustín M.P. 2010. "An Overview of Variables Affecting Lexical Transfer in Writing: A Review Study." *International Journal of Linguistics*, 2(1): E2.

Lohmann, Johannes. 2011. "Do language barriers affect trade?" *Economics Letters*, 110(2): 159–162.

McCallum, John. 1995. "National Borders Matter: Canada-U.S. Regional Trade Patterns." *The American Economic Review*, 85(3): 615–623.

Melitz, Jacques. 2008. "Language and foreign trade." *European Economic Review*, 52(4): 667–699.

Petroni, Filippo, and Maurizio Serva. 2010. "Measures of lexical distance between languages." *Physica A: Statistical Mechanics and its Applications*, 389(11): 2280–2283.

Reher, David, and Miguel Requena. 2009. "The National Immigrant Survey of Spain: A new data source for migration studies in Europe." *Demographic Research*, 20(12): 253–278.

Rose, Andrew K. 2004. "Do We Really Know That the WTO Increases Trade?" *American Economic Review*, 94(1): 98–114.

Serva, Maurizio. 2011. "Phylogeny and geometry of languages from normalized Levenshtein distance." Cornell University Library, Ithaca, NY. `http://arxiv.org/abs/1104.4426v3`.

Swadesh, Morris. 1952. "Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos." *Proceedings of the American Philosophical Society*, 96(4): 452–463.

Tinbergen, Jan. 1962. *Shaping the World Economy: Suggestions for an International Economic Policy.* New York, NY: The Twentieth Century Fund.

Tomz, Michael, Judith L. Goldstein, and Douglas Rivers. 2007. "Do We Really Know That the WTO Increases Trade? Comment." *American Economic Review*, 97(5): 2005–2018.

U.S. Department of Homeland Security. 2010. "Yearbook of Immigration Statistics: 2009." Washington, DC: U.S. Department of Homeland Security, Office of Immigration Statistics. http://www.dhs.gov/xlibrary/assets/statistics/yearbook/2009/ois_yb_2009.pdf.

Tables

Table 1: Closest and Furthest Language Pairs with Respect to the Levenshtein Distance

Closest		Furthest	
Language	Distance	Language	Distance
Distance to English			
Afrikaans	62.08	Vietnamese	104.06
Dutch	63.22	Turkmen	103.84
Norwegian	64.12	Hakka (China)	103.10
Distance to German			
Luxembourgish	42.12	Korean	104.30
Dutch	51.50	Palestinian Arabic	103.72
Westvlaams (Belgium)	57.86	Yoruba (Nigeria)	103.58
Distance to Spanish			
Galician	54.82	Wolof (Senegal)	103.02
Italian	56.51	Igbo Onitsha (Nigeria)	102.84
Portuguese	64.21	Ewondo (Cameroon)	101.87

Notes: – The table shows the three closest and furthest languages towards English, German and Spanish according to the normalized and divided Levenshtein distance. – Only languages spoken within samples are listed. – Geographic origin of language in parentheses.

Table 2: Descriptive Statistics of Dependent and Explanatory Variables – Immigration Sample

	2000 U.S. Census		SOEP		NISS	
	Mean	StdD	Mean	StdD	Mean	StdD
Good language skills	0.58	0.49	0.52	0.50	0.58	0.49
Years of education	11.32	4.28	10.50	2.21	10.82	3.38
Age at entry	26.76	8.72	28.68	8.93	30.00	9.63
Years since migration	12.72	9.91	18.50	11.11	8.33	6.96
Married	0.68	0.47	0.86	0.35	0.58	0.49
One child	0.19	0.39	0.50	0.50	0.24	0.43
Two children	0.19	0.39	0.21	0.41	0.22	0.42
Three or more children	0.14	0.35	0.18	0.38	0.34	0.47
Distance to home country (in 100 km)	57.60	39.95	19.14	14.62	24.41	22.96
Naturalized	0.34	0.47	0.35	0.48	0.07	0.25
Former colony	0.11	0.32	0.10	0.30	0.03	0.18
Southern states	0.29	0.45				
Non-metropolitan area	0.01	0.12				
Minority language share	0.33	0.25				
Abroad five years ago	0.23	0.42				
Refugee	0.12	0.32	0.07	0.25		
Neighboring country			0.12	0.33		
Family abroad			0.30	0.46		
Political reasons					0.03	0.16

Notes: – Number of observations: 59,889 in the 2000 U.S. Census, 675 in the SOEP, and 2,513 in the NISS Sample. – The dependent variable "Good language skills" is defined dichotomously, 1 indicates higher language skills.

Table 3: Immigrant's Language Skills – Probit Results

Dataset:	2000 U.S. Census		SOEP	NISS
Linguistic distance measure:	Test-score	LDND	LDND	LDND
	ME/StdE	ME/StdE	ME/StdE	ME/StdE
Linguistic distance (Test-score-based)	−0.003***			
	(0.000)			
Levenshtein distance (ASJP)		−0.001***	−0.002*	−0.002**
		(0.000)	(0.001)	(0.001)
Years of education	0.048***	0.048***	0.054***	0.029***
	(0.001)	(0.001)	(0.011)	(0.003)
Age at entry	−0.018***	−0.018***	−0.039**	0.006
	(0.002)	(0.002)	(0.015)	(0.007)
Age at entry^2/100	0.012***	0.012***	0.049*	−0.019*
	(0.002)	(0.002)	(0.022)	(0.010)
Years since migration	0.014***	0.014***	0.032**	0.039***
	(0.001)	(0.001)	(0.011)	(0.005)
Years since migration^2/100	−0.021***	−0.021***	−0.058*	−0.067***
	(0.002)	(0.002)	(0.024)	(0.015)
Married	0.020***	0.020***	−0.008	0.003
	(0.005)	(0.005)	(0.066)	(0.026)
Children in the HH. (Ref.= 0)				
One child	0.018**	0.019**	0.013	−0.055†
	(0.006)	(0.006)	(0.083)	(0.030)
Two children	0.015*	0.015*	−0.038	0.072†
	(0.006)	(0.006)	(0.083)	(0.038)
Three or more children	−0.001	−0.000	−0.094	−0.128**
	(0.007)	(0.007)	(0.087)	(0.041)
Distance to home country (in 100 km)	−0.002†	−0.002†	0.006	0.001
	(0.001)	(0.001)	(0.008)	(0.004)
Distance to home country^2/100	0.003***	0.003***	−0.002	0.001
	(0.001)	(0.001)	(0.010)	(0.002)
Naturalized	0.138***	0.138***	0.300***	0.043
	(0.006)	(0.006)	(0.065)	(0.049)
Former colony	0.108***	0.102***	−0.222	0.227***
	(0.011)	(0.011)	(0.218)	(0.059)
Southern states	0.044***	0.045***		
	(0.005)	(0.005)		
Non-metropolitan area	0.051**	0.049**		
	(0.018)	(0.018)		
Minority language share	−0.252***	−0.280***		
	(0.019)	(0.020)		
Abroad five years ago	−0.093***	−0.091***		
	(0.007)	(0.007)		
Refugee	−0.233***	−0.214***	−0.085	
	(0.009)	(0.009)	(0.103)	
Neighboring country			0.310†	
			(0.166)	
Family abroad			−0.069	
			(0.056)	
Political reasons				0.045
				(0.063)
Region fixed effects	yes	yes	yes	yes
Pseudo-R^2	0.263	0.261	0.160	0.138
Observations	59889	59889	675	2513

*Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Robust standard errors are reported in parentheses. – The dependent variable is defined dichotomously, 1 indicates higher language skills. – Probit results are reported as marginal effects evaluated at covariate means. – Region controls are not recorded.*

Table 4: Descriptive Sample Statistics and Variable Definitions – International Trade Sample

	Mean	StdD	Definitions
Log real trade	10.062	3.336	average value of real bilateral trade between countries i and j at year t in US $
Common language	0.223	0.416	binary variable which is unity if i and j share a common language and zero otherwise
Levenshtein distance	90.256	22.453	Levenshtein distance using the most prevalent native language of each country
Levenshtein distance LF	75.063	36.787	Levenshtein distance using the most prevailing lingua franca of each country
Linguistic features index	0.429	0.203	index which increases with decreasing similarity of a language-pair (values between 0 and 1)
Both in GATT/WTO	0.652	0.476	binary variable which is unity if both i and j are GATT/WTO participants at t
One in GATT/WTO	0.307	0.461	binary variable which is unity if either i or j is a GATT/WTO participant at t
General system of preferences	0.231	0.422	binary variable which is unity if i was a GSP beneficiary of j or vice versa at t
Log distance	8.165	0.809	great circle distance between country i and country j in miles
Log product real GDP	47.881	2.676	product of the real GDP's of both countries in year t
Log product real GDP p/c	16.034	1.504	product of the real GDP's per capita of both countries in year t
Regional FTA	0.015	0.120	binary variable which is unity if i and j both belong to the same regional trade agreement
Currency union	0.014	0.118	binary variable which is unity if i and j use the same currency at time t
Land border	0.031	0.172	binary variable which is unity if i and j share a land border
Number landlocked	0.246	0.466	number of landlocked nations in the country-pair (0, 1, or 2)
Number islands	0.341	0.540	number of island nations in the country-pair (0, 1, or 2)
Log product land area	24.206	3.280	product of the land areas of both countries (in square kilometers)
Common colonizer	0.100	0.300	binary variable which is unity if i and j were ever colonies after 1945 with the same colonizer
Currently colonized	0.002	0.044	binary variable which is unity if i is a colony of j at time t or vice versa
Ever colony	0.021	0.142	binary variable which is unity if i ever colonized j or vice versa
Common country	0.000	0.017	binary variable which is unity if i and j remained part of the same nation during the sample

Notes: – Number of observations: 234,597 in 12,150 country-pair groups with 1 to 52 observations per group. The mean is 19.3 observations per group. – For the linguistic features index there are 227,145 observations in 11,348 country-pair groups.

Table 5: Effect of Language on Bilateral Trade – OLS Results

	ComLang Coef/StdE	LDND I Coef/StdE	LDND II Coef/StdE	LingFeat Coef/StdE
Common language	0.274***			
	(0.044)			
Levenshtein distance		−0.006***		
		(0.001)		
Levenshtein distance LF			−0.003***	
			(0.000)	
Linguistic features index				−0.618***
				(0.098)
Both in GATT/WTO	0.604***	0.618***	0.609***	0.578***
	(0.061)	(0.061)	(0.061)	(0.062)
One in GATT/WTO	0.277***	0.288***	0.283***	0.247***
	(0.056)	(0.056)	(0.056)	(0.056)
General system of preferences	0.709***	0.733***	0.711***	0.721***
	(0.032)	(0.031)	(0.032)	(0.032)
Log distance	−1.313***	−1.278***	−1.308***	−1.293***
	(0.023)	(0.024)	(0.023)	(0.024)
Log product real GDP	0.167**	0.165**	0.164**	0.159**
	(0.051)	(0.051)	(0.051)	(0.053)
Log product real GDP p/c	0.532***	0.533***	0.535***	0.552***
	(0.049)	(0.049)	(0.049)	(0.050)
Regional FTA	0.941***	0.942***	0.939***	0.975***
	(0.126)	(0.126)	(0.126)	(0.129)
Currency union	1.174***	1.253***	1.169***	1.208***
	(0.122)	(0.125)	(0.123)	(0.124)
Land border	0.280**	0.283**	0.284**	0.292**
	(0.108)	(0.108)	(0.108)	(0.113)
Number landlocked	−1.056***	−1.032***	−1.014***	−0.971***
	(0.207)	(0.205)	(0.205)	(0.208)
Number islands	−1.579***	−1.575***	−1.622***	−1.545***
	(0.188)	(0.188)	(0.188)	(0.190)
Log product land area	0.496***	0.501***	0.513***	0.496***
	(0.041)	(0.041)	(0.041)	(0.041)
Common colonizer	0.605***	0.703***	0.592***	0.687***
	(0.064)	(0.062)	(0.065)	(0.065)
Currently colonized	0.743**	0.744**	0.753**	0.719**
	(0.263)	(0.252)	(0.264)	(0.262)
Ever colony	1.274***	1.272***	1.261***	1.339***
	(0.114)	(0.116)	(0.114)	(0.113)
Common country	0.288	0.090	0.263	0.278
	(0.583)	(0.658)	(0.579)	(0.617)
Year fixed effects	yes	yes	yes	yes
Country fixed effects	yes	yes	yes	yes
Adjusted R^2	0.703	0.703	0.703	0.705
RMSE	1.818	1.817	1.817	1.805
F Statistic	274.34***	272.11***	274.96***	265.50***
Observations	234597	234597	234597	227145

*Notes: – Significant at: *** 0.1% level; ** 1% level; * 5% level; † 10% level. – Robust standard errors (clustered at the country-pair lvel) are reported in parentheses. – The dependent variable is defined as log of real bilateral trade in US$. – Intercept, year, and country controls are not recorded.*

Appendix

Table A1: 40-Items Swadesh Word List

I	You	We	One
Two	Person	Fish	Dog
Louse	Tree	Leaf	Skin
Blood	Bone	Horn	Ear
Eye	Nose	Tooth	Tongue
Knee	Hand	Breast	Liver
Drink	See	Hear	Die
Come	Sun	Star	Water
Stone	Fire	Path	Mountain
Night	Full	New	Name

Source: Bakker et al. (2009).

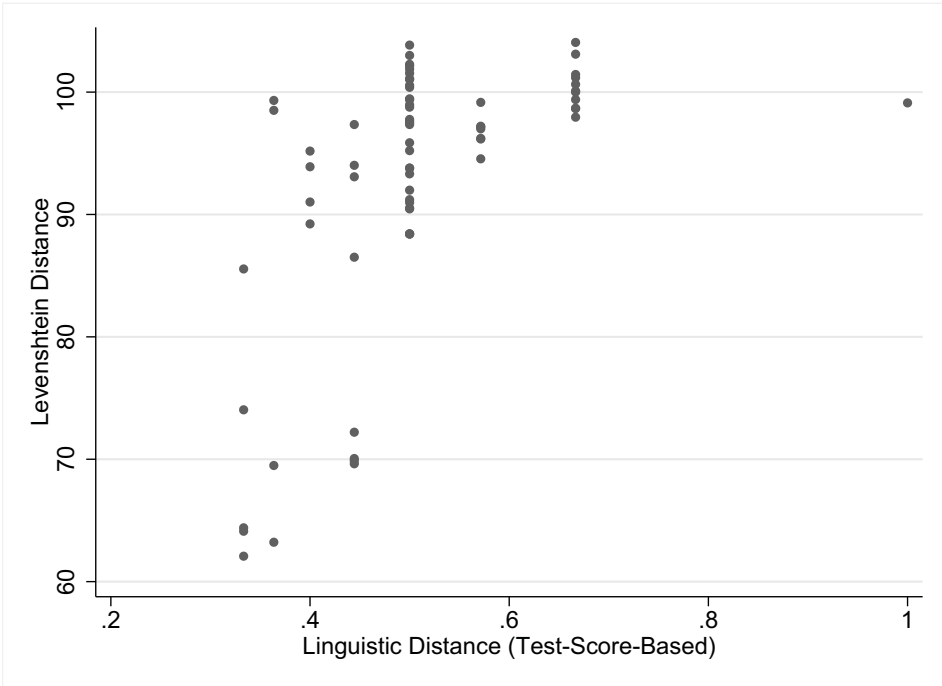

Figure A1: Comparisons of Linguistic Distance Using the Test-Score-Based Measure and the Levenshtein Distance – 2000 U.S. Census

Table A2: Summary Statistics For The Language Variables – International Trade Sample

A. Simple Correlations among Language Distance Measures

	Common language	Levenshtein distance	Levenshtein distance LF	Linguistic features index[a]
Common language	1			
Levenshtein distance	-0.3868	1		
Levenshtein distance LF	-0.6689	0.4813	1	
Linguistic features index[a]	-0.3533	0.5490	0.4070	1

B. Frequency of Country-pairs with and without the same Language

	Common language	Levenshtein distance	Levenshtein distance LF	Linguistic features index[a]
Same language	52,205	11,017	43,229	21,389
Different language	182,392	223,580	191,368	205,756

Notes: – Number of observations: 234,597, except [a] 227,145.

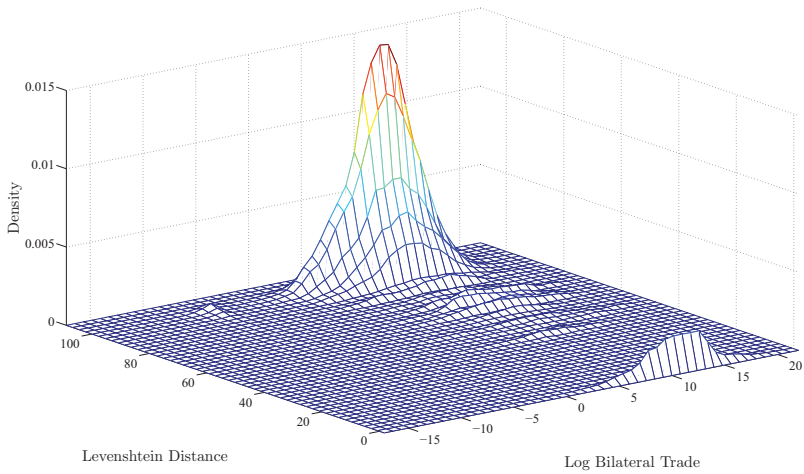

Figure A2: Bivariate Kernel Density Estimation of Log Bilateral Trade and Levenshtein Distance – International Trade Sample

Sensitivity Analysis

As mentioned above, we perform a number of sensitivity analyses which, in each case, find similar results to those reported above. First, we estimate our model of immigrant's language skills using an expanded sample which also includes individuals speaking the language of the host country as mother tongue. The results are reported in Table A3.

Second, we repeat our estimations using the four or fivefold information of language fluency as dependent variable in Ordered Probit models. The marginal effects across the different categories indicate a comparable effect as in the dichotomous case. The signs of the marginal effects change at the same threshold we use for the dichotomization. The results are available upon request.

Third, Tables A4 and A5 report estimation results for two subsamples of the trade sample. The results are quite similar in magnitude and significance level to those for the whole sample. Table A4 examines the sensitivity of the results with respect to the measurement of linguistic distance. Therefore, we exclude dyad-observations with the same language from the sample, thereby including only country-pairs with a language barrier greater than zero. This tests the idea that country-pairs speaking or not speaking a common language delivering the results of the language barrier, rather than an effect of linguistic distance *per se*. Table A5 analyzes the sensitivity of our results when we restrict our sample to the slightly smaller one of Lohmann's linguistic features index.

Finally, we add for both countries in a country-pair country-by-time interaction terms, $\sum_{it} \eta_{it} (I \times T)_{it}$ and $\sum_{jt} \psi_{jt} (J \times T)_{jt}$, to the models of Table 5. These interaction terms capture any exporter and importer specific time-variant effects such as each country's business cycle or its institutional characteristics. The findings for the key variables (available upon request) are quite similar in magnitude and significance level to those for the models with country and year specific fixed effects.

Table A3: Immigrant's Language Skills – Probit Results, Including Native Speakers

	ME/StdE	ME/StdE	ME/StdE	ME/StdE
Linguistic distance (Test-score-based)	−0.008***			
	(0.000)			
Levenshtein distance (ASJP)		−0.007***	−0.003*	−0.004***
		(0.000)	(0.001)	(0.000)
Years of education	0.040***	0.040***	0.054***	0.015***
	(0.001)	(0.001)	(0.011)	(0.002)
Age at entry	−0.014***	−0.015***	−0.038*	0.002
	(0.001)	(0.001)	(0.015)	(0.003)
Age at entry^2/100	0.009***	0.010***	0.047*	−0.008
	(0.002)	(0.002)	(0.021)	(0.005)
Years since migration	0.013***	0.013***	0.029**	0.022***
	(0.001)	(0.001)	(0.011)	(0.003)
Years since migration^2/100	−0.020***	−0.022***	−0.053*	−0.038***
	(0.002)	(0.002)	(0.024)	(0.008)
Married	0.015**	0.018***	0.007	0.003
	(0.005)	(0.005)	(0.066)	(0.014)
Children in the HH. (Ref.= 0)				
One child	0.012*	0.016**	0.008	−0.030†
	(0.005)	(0.005)	(0.082)	(0.017)
Two children	0.012*	0.014*	−0.039	0.035*
	(0.005)	(0.005)	(0.082)	(0.018)
Three or more children	−0.005	0.001	−0.092	−0.065**
	(0.006)	(0.006)	(0.086)	(0.022)
Distance to home country (in 100 km)	−0.000	0.000	0.007	0.001
	(0.001)	(0.001)	(0.008)	(0.002)
Distance to home country^2/100	0.002***	0.002***	−0.003	0.000
	(0.000)	(0.000)	(0.010)	(0.001)
Naturalized	0.108***	0.106***	0.268***	0.017
	(0.004)	(0.005)	(0.066)	(0.024)
Former colony	0.123***	0.095***	−0.203	0.240***
	(0.008)	(0.008)	(0.202)	(0.030)
Southern states	0.049***	0.064***		
	(0.004)	(0.004)		
Non-metropolitan area	0.039**	0.021		
	(0.015)	(0.015)		
Minority language share	−0.340***	−0.567***		
	(0.016)	(0.017)		
Abroad five years ago	−0.084***	−0.078***		
	(0.006)	(0.006)		
Refugee	−0.266***	−0.214***	−0.096	
	(0.009)	(0.008)	(0.102)	
Neighboring country			0.313*	
			(0.159)	
Family abroad			−0.085	
			(0.056)	
Political reasons				0.032
				(0.029)
Region fixed effects	yes	yes	yes	yes
Pseudo-R^2	0.304	0.299	0.165	0.347
Observations	70201	70201	689	3986

*Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Robust standard errors are reported in parentheses. – The dependent variable is defined dichotomously, 1 indicates higher language skills. – Probit results are reported as marginal effects evaluated at covariate means. – Region controls are not recorded.*

Table A4: Effect of Language on Bilateral Trade – OLS Results, Subsample Language Barrier > 0

	LDND I Coef/StdE	LDND II Coef/StdE	LingFeat Coef/StdE
Levenshtein distance	-0.008^{***}		
	(0.002)		
Levenshtein distance LF		-0.008^{***}	
		(0.002)	
Linguistic features index			-0.266^{*}
			(0.123)
Both in GATT/WTO	0.506^{***}	0.435^{***}	0.454^{***}
	(0.067)	(0.070)	(0.069)
One in GATT/WTO	0.219^{***}	0.193^{**}	0.174^{**}
	(0.061)	(0.064)	(0.063)
General system of preferences	0.754^{***}	0.612^{***}	0.682^{***}
	(0.031)	(0.032)	(0.032)
Log distance	-1.281^{***}	-1.211^{***}	-1.271^{***}
	(0.025)	(0.027)	(0.027)
Log product real GDP	0.060	-0.070	-0.005
	(0.053)	(0.058)	(0.056)
Log product real GDP p/c	0.646^{***}	0.793^{***}	0.727^{***}
	(0.051)	(0.056)	(0.054)
Regional FTA	0.828^{***}	-0.252^{*}	0.190
	(0.147)	(0.118)	(0.155)
Currency union	1.312^{***}	1.138^{***}	1.200^{***}
	(0.133)	(0.275)	(0.203)
Land border	0.281^{*}	0.391^{**}	0.301^{*}
	(0.120)	(0.130)	(0.133)
Number landlocked	-1.232^{***}	-1.330^{***}	-1.307^{***}
	(0.206)	(0.207)	(0.214)
Number islands	-1.837^{***}	-2.492^{***}	-1.993^{***}
	(0.192)	(0.199)	(0.203)
Log product land area	0.551^{***}	0.668^{***}	0.581^{***}
	(0.042)	(0.043)	(0.044)
Common colonizer	0.697^{***}	0.887^{***}	0.662^{***}
	(0.063)	(0.092)	(0.069)
Currently colonized	0.322	1.149^{**}	0.442
	(0.292)	(0.391)	(0.390)
Ever colony	1.517^{***}	1.041^{***}	1.182^{***}
	(0.131)	(0.193)	(0.152)
Common country	1.203^{***}		
	(0.346)		
Year fixed effects	yes	yes	yes
Country fixed effects	yes	yes	yes
Adjusted R^2	0.702	0.706	0.708
RMSE	1.827	1.792	1.801
F Statistic	860.28^{***}	267.50^{***}	275.73^{***}
Observations	223580	191368	205756

*Notes: – Significant at: *** 0.1% level; ** 1% level; * 5% level; † 10% level. – Robust standard errors (clustered at the country-pair level) are reported in parentheses. – The dependent variable is defined as log of real bilateral trade in US$. – Intercept, year, and country controls are not recorded. – In column (2) and (3) common country is omitted from the equations because of collinearity.*

Table A5: Effect of Language on Bilateral Trade – OLS Results, Subsample Linguistic Features Index

	ComLang Coef/StdE	LDND I Coef/StdE	LDND II Coef/StdE
Common language	0.292***		
	(0.045)		
Levenshtein distance		−0.006***	
		(0.001)	
Levenshtein distance LF			−0.004***
			(0.001)
Both in GATT/WTO	0.585***	0.600***	0.592***
	(0.062)	(0.062)	(0.062)
One in GATT/WTO	0.255***	0.267***	0.264***
	(0.057)	(0.056)	(0.057)
General system of preferences	0.707***	0.732***	0.708***
	(0.032)	(0.031)	(0.032)
Log distance	−1.305***	−1.268***	−1.297***
	(0.023)	(0.024)	(0.023)
Log product real GDP	0.166**	0.163**	0.164**
	(0.053)	(0.053)	(0.052)
Log product real GDP p/c	0.546***	0.547***	0.548***
	(0.050)	(0.050)	(0.050)
Regional FTA	0.980***	0.980***	0.975***
	(0.129)	(0.128)	(0.128)
Currency union	1.185***	1.268***	1.172***
	(0.123)	(0.126)	(0.124)
Land border	0.285*	0.290**	0.293**
	(0.113)	(0.112)	(0.112)
Number landlocked	−1.033***	−1.009***	−0.987***
	(0.209)	(0.207)	(0.207)
Number islands	−1.602***	−1.599***	−1.655***
	(0.191)	(0.190)	(0.190)
Log product land area	0.493***	0.498***	0.510***
	(0.042)	(0.041)	(0.041)
Common colonizer	0.595***	0.701***	0.570***
	(0.067)	(0.065)	(0.068)
Currently colonized	0.734**	0.734**	0.744**
	(0.268)	(0.256)	(0.269)
Ever colony	1.255***	1.251***	1.222***
	(0.115)	(0.117)	(0.114)
Common country	0.307	0.103	0.281
	(0.596)	(0.678)	(0.592)
Year fixed effects	yes	yes	yes
Country fixed effects	yes	yes	yes
Adjusted R^2	0.705	0.706	0.706
RMSE	1.805	1.804	1.805
F Statistic	268.30***	265.73***	269.23***
Observations	227145	227145	227145

*Notes: – Significant at: *** 0.1% level; ** 1% level; * 5% level; † 10% level. – Robust standard errors (clustering at the country-pair level) are reported in parentheses. – The dependent variable is defined as log of real bilateral trade in US$. – Intercept, year, and country controls are not recorded.*