The Costs of Babylon – Linguistic Distance in Applied Economics*

Ingo E. Isphording[a,b] and **Sebastian Otten**[a,c]

[a]Ruhr-University Bochum
[b]RUB Research School
[c]RWI Essen

This version: October 2011

Abstract

Linguistic distance, i.e. the dissimilarity between languages, is an important factor influencing international economic transactions such as migration or international trade flows. Linguistic distance imposes costs for acquiring a second language and, higher linguistic distance between countries rises hurdles for e.g. language acquisition of immigrants, or induces higher costs for international transactions. We introduce a new measure of linguistic distance based on recent research by the German Max Planck Institute of Evolutionary Anthropology. 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, the linguistic distance reduces reported language skills of immigrants. Second, using data of international trade flows covering 175 countries and 50 years in a standard gravity model, it is shown that linguistic distance has a strong negative influence on bilateral trade volumes.

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

JEL classifications: J24, J61, F22, F16

*The authors are grateful to Thomas K. Bauer and John P. Haisken-DeNew 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-University Bochum, 44780 Bochum, Germany, Email: 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, the linguistic diversity and linguistic distance between languages has been shown to be a crucial determinant of real economic outcomes, due to its impact on communication and language skills, and has been used in analyzes of immigrant language skills and labor market integration (see, e.g., Chiswick and Miller 1999, 2005, 2011). This influence on the micro level might accumulate to substantial costs on the macro level, by affecting international trade flows (see, e.g., Hutchinson 2005, Lohmann 2011).

The operationalization of linguistic distance is quite complicated. Previous studies usually relied on approaches measuring linguistic distance using average test scores of language students (Chiswick and Miller 1999). This approach assumes the difficulty of learning a foreign language for students to be determined by the distance between the native and foreign language. Unfortunately, so far the test-score-based measures are only available for the distances towards the English language, and are therefore strongly restricted in its use. Other studies use a language-tree approach based on classifications by language trees (Guiso, Sapienza and Zingales 2009).

To apply a more comprehensive and advantageous measure that accounts for differences in all of the world's languages, this study proposes to use an easily and transparently computed continuous measure. It was originally developed by the German Max Planck Institute of Evolutionary Anthropology to explain geographical diversity of languages. The purely descriptive measure of phonetic similarity is based on the automatic comparison of the pronunciation of words from different languages having the same meaning. A first application of this measure in the setting of language acquisition of immigrants can be found in Isphording and Otten (2011).

We show the merits of this measure exemplarily in two different potential applications. First, we apply the measure to analyze its explanatory power in the case of language acquisition of migrants. Linguistic distance is assumed to increase the costs and lower the efficiency of second language acquisition, thereby ceteris paribus decreasing language proficiency in the host country language. We use data from the 2000 U.S. Census, the German Socio-Economic Panel, and the National Immigrant Survey of Spain in order to be able to draw conclusions from an international comparison, and to directly compare the Levenshtein distance to the previously used test-score-based measure of linguistic distance by Chiswick and Miller (1999).

Second, we apply the measure in the context of international trade. Linguistic bar-

riers are likely to affect international trade flows by imposing higher transaction costs, e.g. by increased costs of translation. This effect has been addressed previously by simply including indicator variables controlling for a shared language between trade partners (Anderson and van Wincoop 2004). To overcome this very narrow definition of linguistic barriers, we apply the Levenshtein distance by utilizing a comprehensive dataset on bilateral trade flows by Rose (2004), and compare it with an alternative approach by Lohmann (2011). Our results indicate that not only a shared common language accelerates trade, but also related but not identical languages have an increasing effect on trade by lowering transaction costs.

The paper is organized as follows. We begin with a short overview on previous attempts of measuring linguistic distance and then introduce the Levenshtein distance, discussing its advantages and potential shortcomings in Section 2. We present our results obtained by applying this approach for 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 multidimensionality renders it difficult to come to a specific definition and an appropriate empirical representation to be used in applied economic studies.

A widely used approach has been introduced by Chiswick and Miller (1999), who use data on the average test score of U.S. American 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. The measure has been successfully applied in migration economics (Chiswick and Miller 1999, 2005, 2011) and international trade (Hutchinson 2005). Following this idea, Ku and Zussman (2010) have used TOEFL test data to compute a similar measure.

Although it is only available for the distance to English, this measure allows for a comprehensive comparison of languages across different dimensions, yet it has to rely on strong assumptions. To use this measure to represent linguistic distance, it has to be assumed that the difficulty for U.S. Americans to learn the 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 on potential intrinsic and extrinsic motivation to learn a second language. Intrinsic motivation, the inherent pleasure of learning a language, and extrinsic motivation, the utility one derives from being able to communicate in the foreign language, both most likely differ across languages, but are not distinguishable from the actual linguistic distance in the test-score-based approach.

Other approaches rely on the historical development of languages to separate them into broader families, and to define certain distances between languages. Lehmann (1992) offers an introduction to these historical linguistics. Such a measure has already been used in the management literature by West and Graham (2004) to explain differences in managerial values, and in the economic literature by Guiso, Sapienza and Zingales (2009) to explain international trade volumes. But although encompassing all different dimensions of language, the grouping within family trees according to the historical development leads to a measure with only very few increments and little variation.

2.2 The Levenshtein Distance

The Levenshtein distance as a measure of distance between different character strings was introduced by Levenshtein (1966). The linguistic interpretation we are going to rely on is based on an approach developed by the German Max Planck Institute for Evolutionary Anthropology. The so-called *Automatic Similarity Judgement Program* (ASJP) 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 (Brown et al. 2008).

This distance gives an approximation of the number of cognates between languages. Cognates is a linguistic term which denotes common ancestries of words. A higher number of cognates indicates closer common ancestries. Therefore, a lower Levenshtein distance also indicates a higher probability of sharing other language characteristics such as grammar, while focusing in its computation on differences in pronunciation (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 (Kuhl and Iverson 1995, Best, McRoberts and Goodell 2001).

The algorithm judging the distance between words relies on a specific phonetic alphabet, the ASJPcode. The ASJPcode uses all available characters within the standard ASCII 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 for how many sounds have to be substituted, added, or removed to transfer the one word into the other (Holman et al. 2011). The words used in the approach are taken from the so-called 40-item Swadesh list, which is a 40-word list including words which 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.[1]

The ASJP program judges each word pair across languages according their similarity in pronunciation. Example: To transfer the phonetic transcription of the English word *you*, *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*, one has to remove or substitute each consonant and vowel.

To account for differences in word length, the resulting number of changes is divided by the word length of the longer one. To additionally account for the number of word pairs that exist between two languages i and j, the normalized and divided linguistic distance (LDND) is calculated as:

$$LDND = \frac{\sum_i (d_{ii})/n}{\sum_{i \neq j} (d_{ij})/n(n-1)}, \tag{1}$$

where d_{ii} is the distance between item i in language A and item i in language B, hence between words of the same meaning. d_{ij} denotes the distance between words with different meanings, which accounts for similarities by chance in phonetic inventories. n denotes the number of existing word pairs between languages.

Table 1 lists the closest and furthest languages with respect to English. The measurement via the ASJP approach seems to be in line with intuitive guessing about language dissimilarity towards English. Figure A1 in the Appendix shows the relationship between the test-score-based approach by Chiswick and Miller (1999) and the ASJP approach. Although there is clearly a strong positive correlation, the ASJP offers a higher variability in its measurement. Some languages are found to be distant by the ASJP measure, but have a comparably low distance by using the test-score-based measure, indicating that the test-score-based measure might also entail incentives to learn a foreign language, instead solely measuring linguistic distance.

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

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 starts off with the influential work by Chiswick (1991). In following work, Chiswick and Miller (1995) developed a theoretical human capital framework of host country language skill acquisition. In this framework, the 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 is also consistent with the sociological rational choice model by Esser (2006), and has been subsequently tested for various countries using the test scores-based measure of Chiswick and Miller (1999). Due to its only availability to the English language, these applications have been restricted to studies concerning the immigration to English-speaking countries like the United States 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 makes it feasible to broaden the evidence on the relationship between linguistic distance and language fluency to an international perspective. To do this, 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 within in the same dataset. To compare the influence across different countries, we additionally use German data from the Socio-Economic Panel, and the National Immigrant Survey of Spain.

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 about 1 million immigrants are granted legal permanent resident status per year. In 2000, this immigration consisted mainly of immigrants from other North-American countries (40 %, 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, 11.1 % of the population of the United States were foreign-born. The immigrant population is considerably younger than the native population and dominated by low-skilled immigrants. Among those migrated 1990 and later, 34.4 % do not possess a high-school degree. This share has been risen from 19.3 % before 1970. Within the native population, only 8 % possessed no high-school degree in 2000 (Camarota 2001).

Germany does not have such a long-running immigration history as the U.S., neither can it look back on an extensive colonial history as Spain. Mass immigration only

started off shortly after World War II, with the so-called "Guestworker"-programs to attract mainly unskilled workers from Mediterranean countries such as Turkey, Yugoslavia, Italy, or Spain. This first wave of immigration was 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. In 2009, 10.6 million (approx. 13 %) of the German population have immigrated after 1949, 3.3 million as Ethnic Germans. The major part is originated in EU member states (32.2 %), followed by 28 % from Turkey and 27 % from former members of the Soviet Union. Compared to the U.S. and Spain, Germany has a very old immigrant population, with long individual migration histories. The immigrants are considerably lower educated than the native population: 14.0 % possess no educational degree (Statistisches Bundesamt 2010).

Although Spain has a long-running colonial history, it is a comparably young immigration country. After large waves of emigration until the 1970s, incoming migration started during the last decades of the last century, and accelerated considerably during the last 20 years. Between 1997 and 2007, the number of migrants increased by approx. 700 %, first including mostly migrants from Latin America, Africa and Western Europe, and since the EU enlargement increasingly from Eastern Europe. Nowadays, about 10 % or 4,5 million individuals are foreign-born, mostly immigrated from Europe, Latin America, and Africa (see, Fernández and Ortega 2008, Amuedo-Dorantes and De la Rica 2007).

3.1 Data and Method

Our data is restricted to male immigrants who entered the respective country after the age of 16 and who younger than 65, and who do not speak the host country language as first language. The sample from the 1% - PUMS (Public Use Microdata Series) 2000 U.S. Census file consists of 59,890 individuals. Similar data is extracted from the German Socio-Economic Panel, a long-run longitudinal reprasentative study. Using cross-sectional data from 2001, the sample consists of 698 male immigrants.[2] The National Immigrant Survey of Spain, conducted in 2007, also offers comprehensive cross-sectional information on the socio-economic characteristics and migration history to immigrants.[3] The sample includes 2,540 male immigrants.

All datasets include self-reported assessments of language fluency in four- or fivefold measures, which are recoded into dichotomous measures, where one means "Good" or "Very Good" language skills as opposed to all lower values. This variable serves as a the

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

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

dependent variable in Probit regressions. The explanatory variables are chosen to ensure the highest possible comparability between the regressions. The linguistic distance enters the specifications 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, and number of children. We additionally include distance in kilometers between home and host country 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 minority language share. A dummy for being originated in a former colony is included in the U.S. and the Spanish regressions. For Germany, we include a dummy for being originated in a neighboring country. We control for whether somebody entered the host country as a refugee (U.S. and Germany) or somebody migrated due to political reasons (Spain). The U.S. data includes further information somebody whether one was abroad 5 years ago, and the German data includes information on having family abroad.

Sample means of the used variables reported in Table 2 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 has also the lowest mean education, but has a higher share of married couples and a higher number of children, which is partly 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 Spain and the United States have a higher 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 of in the 1990s.

The share of individuals reporting "Good" or "Very Good" language skills in the host country language are similar of around slightly more than the half of the sample.

3.2 Results

Table 3 lists the results of the Probit regressions across datasets. 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 prob-

ability 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 ASJP measure are qualitatively comparable, the effect is lower by one third when applying the ASJP measure.

The magnitude of the effect can be illustrated in a distributional context. In the U.S., an individual coming from a linguistically close country in the first percentile of the Levenshtein distance distribution has a higher probability of 20 % of reporting high language skills than an individual coming from a distant country in the highest percentile. This is comparable with 4 additional years of education, all other things equal. In Germany, the probability is 40 % higher, comparable to even approx. 7 additional years of schooling. In Spain, moving up from the bottom to the top of the distribution of linguistic distance still increases the probability of reporting good language skills by 20 %, due to the lower effect of education comparable to 6 additional years of education. These examples assume the marginal effect to be linear across the range of linguistic distance, which has been shown to be true for the German case by Isphording and Otten (2011).

Switching the measure of linguistic distance in the U.S. data does not affect qualitatively the coefficients of the control variables. The coefficients are in line with previous studies and theoretical predictions. We see a positive impact of education, and a negative effect of age at migration, but with a decreasing rate. Years since migration increase the probability of reported language skills, again with decreasing rate. Being married and having a higher number of children also increases the language skills. These qualitative relationships are stable across all datasets, with exception of lower language skills for immigrants with children in Spain. Being originated 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.

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. Anderson and van Wincoop (2004) report in their survey article 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.[4] The question is whether and how much the dissimilarity between two languages matters if trading partners not share a common language?

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 sizes, measured as (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 include language-related barriers in these gravity models, common empirical practice is to use an indicator variable that is one if two countries share the same official language and zero 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 which is one 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. This degree of similarity 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). This is also indicated by the results of section 3.2, where we have shown that linguistic barriers might crucially affect second language acquisition. Further, 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 into account similarities and differences between a multitude of languages are the ones by Hutchinson (2005) and Lohmann (2011). Hutchinson (2005) is restricted to distances towards English by relying on the measure by Chiswick and Miller (1999).

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 lan-

[4]The tax equivalent associated with language barriers reported by Anderson and van Wincoop (2004) is based on estimates of Hummels (1999) and Eaton and Kortum (2002).

guage features within language pairs and builds up a language features index normalized to the interval of $[0; 1]$, where 0 means sharing all features.

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 and Zussman (2010), and Lohmann (2011).[5] The sample covers bilateral trade between 175 countries over the years 1948 to 1999 leading to 234,597 dyad-year observations.

Descriptive statistics of the variables used in the empirical analysis are shown in Table 4. The variables of interest are Rose's binary common language variable, both versions of linguistic distance between trading partners languages as measured by the Levenshtein distance, and finally the linguistic features index calculated by Lohmann. 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 already came 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 United States 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.

The Levenshtein distance is computed for every country-pair in the dataset. In multilingual 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.[6] To analyze the sensitivity of the results with respect to the measurement of the linguistic distance, we calculate an alternative specification of the Levenshtein distance, replacing the most prevalent language with the prevailing lingua franca in a country. We compare the effect of these two definitions of the Levenshtein distance with the approach by Lohmann (2011).

We use the gravity model to estimate the impact of language barriers on trade between

[5]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).

[6]A comprehensive index of assigned languages with further explanations is available upon request.

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 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 used below is:

$$\begin{aligned} ln\, X_{ijt} = \beta_0 + \beta_1\, ln\,(Y_i Y_j)_t + \beta_2\, ln\, Dist_{ij} + \beta_3\, Z_{ijt} + \gamma_1\, LangBar_{ij} \\ + \sum\nolimits_i \delta_i I_i + \sum\nolimits_j \theta_j J_j + \sum\nolimits_t \phi_t T_t + \varepsilon_{ijt}, \end{aligned} \tag{2}$$

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 kilometers. Z is a matrix of control variables, including the population, geographic characteristics like 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 participation in the GATT/WTO (one or both countries), same currency, regional trade agreements, and being a GSP beneficiary.[7]

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 year and country fixed effects accounts for common shocks and trends (e.g., global business cycle and the decline in communication costs) and country-specific effects. The gravity model is estimated by ordinary least squares (OLS) with robust standard errors clustered on country-pairs.

[7]More details are given in Rose (2004) and Tomz, Goldstein and Rivers (2007), the sources for all variables except the linguistic distance measures.

4.2 Results

Table 5 summarizes the results of Eq. (2). 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 $(exp(0.265)-1 \approx)$ 30.3 %. 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. For an easier interpretation of the effect of the Levenshtein distance, the elasticity and the marginal effect multiplied by the interquartile range of the Levenshtein distance are calculated. Increasing the Levenshtein distance by 1 % decreases bilateral trade by about 0.5 %. Moving up the distribution of the Levenshtein distance from the lower to the upper quartile decreases trade between countries by $(exp(-0.042) - 1 \approx)$ 4.1 %.[8]

In multi-lingual countries, the assignment of languages to countries is difficult. To show that our findings are not a result of a peculiar 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

[8]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 become even lager in magnitude and are stable in significance, while the linguistic features index becomes distinctly smaller in magnitude and significance.

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 only computable for a subsample of 227,145 county-pairs.[9] The coefficient reveals that a pair of countries trade about $(exp(-0.606)-1 \approx)$ 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) and 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 %. Moving up the distribution of the linguistic features index from the lower to the upper quartile decreases trade between countries by $(exp(-1.435) - 1 \approx)$ 7.6 %. 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 analyzis provides evidence that both measures have a statistically and economically significant negative effect on bilateral trade flows.

The control variables in all models act as in the previous literature. The indicators for whether one or both countries in the dyad participated in the GATT/WTO have negative coefficients. However, we only find significantly less trade when one country of the trading partners has GATT/WTO rights and obligations compared to country-pairs in which neither country belongs to the agreement. The respective coefficients are even more negative and statistically significant as compared to those reported by Rose (2004) and Tomz, Goldstein and Rivers (2007). A plausible explanation for this result is that our model differs from those of Rose (2004) and Tomz, Goldstein and Rivers (2007) in two ways: Compared to Rose, we use a corrected sample while we amend the approach of Tomz et al. by additionally including country fixed effects. Apart from the GATT/WTO membership variables the model confirms the traditional results of gravity trade equations. 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

[9]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.

the literature.

As compared to the first specification, the application of the Levenshtein distance measure does not affect the magnitude or significance of the other independent variables considerably. 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. Since colonizers often change the legal and administrative system in their colonies, the use of the colonizer language in official communication was common during colonization times and even after the colonial era. Therefore, some knowledge in the colonizer's language is not only common for the older but also for the younger population, e.g. it might be the first foreign language children learn at school. Thus, a common colonizer promotes trade between these countries, which might be hidden when not controlling properly for linguistic heterogeneity.

5 Conclusion

This study introduced a new easily and transparently computed measure of linguistic distance, and showed its potential applications in applied economics. The measure relies on linguistic research from 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 for 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 of measuring linguistic distance by Chiswick and Miller (1999) 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 Levenshtein distance from the ASJP data offers a simple and comprehensive way of including linguistic distance into applied economic studies that have the need to control for international differences in linguistic origin. It avoids potential omitted variable biases or biases by unobserved heterogeneity when linguistic distance is ignored. The merits of the Levenshtein distance were shown in two applications, the language acquisition of immigrants, and the analysis of bilateral trade flows.

Following the widely accepted rational choice framework of language acquisition (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 micro data. The results confirm a strong significantly negative effect of linguistic distance on immigrant language skills. Moving up the distribution of linguistic distance from the lowest to highest quartile decreases the probability of reporting good language skills by up to 40 %. As such, the linguistic distance is able to explain a large part of language skill heterogeneity in immigrant populations.

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 data set 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). Our 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 to the upper quartile decreases trade between countries by about 4.1 %.

Taken together, this study showed the significant economic costs of linguistic heterogeneity on the micro and the macro level, the "costs of Babylon", referring to biblical accounts of the Babylonian Confusion. The Levenshtein distance offers an 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

Amuedo-Dorantes, Catalina, and Sara De la Rica. 2007. "Labour Market Assimilation of Recent Immigrants in Spain." *British Journal of Industrial Relations*, 45(2): 257–284.

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.

Best, Catherine T., Gerald W. McRoberts, and Elizabeth Goodell. 2001. "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system." *The Journal of the Acoustical Society of America*, 109(2): 775–794.

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 2008. "Automated classification of the world/s languages: a description of the method and preliminary results." *STUF – Language Typology and Universals*, 61(4): 285–308.

Camarota, Steven A. 2001. "Immigrants in the United States – 2000: A Snapshot of America's Foreign-Born Population." Center for Immigration Studies. `http://cis.org/articles/2001/back101.pdf`.

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.

Chiswick, Barry R., and Paul W. Miller. 2011. "Negative and Positive Assimilation, Skill Transferability, and Linguistic Distance." Institute for the Study of Labor (IZA) Discussion Paper No. 5420.

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

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

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica*, 70(5): 1741–1779.

Esser, Hartmut. 2006. "Migration, Language and Integration: AKI Research Review 4." Berlin: Social Science Research Center 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/education/languages/Focus/docs/elan_en.pdf`.

Haisken-DeNew, John P., and Joachim R. Frick. 2005. "Desktop Companion to the German Socio-Economic Panel (SOEP): Version 8.0." Berlin: German Institute for Economic Research. `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." `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*, forthcoming.

Hummels, David. 1999. "Toward a Geography of Trade Costs." GTAP Working Papers No. 17.

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.

Kuhl, Patricia K., and Paul Iverson. 1995. "Linguistic experience and the "perceptual magnet effect"." In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research.* ed. Winifred Strange, 121–154. Baltimore and MD: York Press.

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

Lehmann, Winfred P. 1992. *Historical Linguistics: An Introduction.* 3rd ed. London: Routledge.

Levenshtein, Vladimir I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals." *Soviet Physics Doklady*, 10(8): 707–710.

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.

Reher, David, and Miguel Requena. 2009. "The National Immigrant Survey of Spain: A new data source for migration studies in Europe." *Demographic Research*, 20: 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." `http://arxiv.org/abs/1104.4426v3`.

Statistisches Bundesamt. 2010. "Bevölkerung mit Migrationshintergrund – Ergebnisse des Mikrozensus 2009." Wiesbaden: Statistisches Bundesamt. http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Publikationen/Fachveroeffentlichungen/Bevoelkerung/MigrationIntegration/Migrationshintergrund2010220097004,property=file.pdf.

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: 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 and D.C.: U.S. Department of Homeland Security, Office of Immigration Statistics.

West, Joel, and John L. Graham. 2004. "A Linguistic-based Measure of Cultural Distance and Its Relationship to Managerial Values." *MIR: Management International Review*, 44(3): 239–260.

Tables

Table 1: Closest and Furthest Immigrant Languages with Respect to Levenshtein Distance to English

Closest		Furthest	
Language	*Distance*	*Language*	*Distance*
Afrikaans	62.08	Vietnamese	104.06
Dutch	63.22	Turkmen	103.84
Norwegian	64.12	Hakka[a]	103.10
Swedish	64.40	Cambodian	103.00
Frisian	69.49	Finnish	102.27

Notes: – The Levenshtein distances are calculated based on immigrants in the 2000 U.S. Census. There exists several closer and further languages to English, but they are not included in the 2000 U.S. Census. For example, in the trade sample Vincentian Creole English (= 41.57) spoken in Saint Vincent and the Grenadines is the closest language to English. The furthest language is Vietnamese as well. – [a] Hakka is a form of Chinese.

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

	2000 U.S. Census Mean/StdD	SOEP Mean/StdD	NISS Mean/StdD
Good language skills	0.58	0.53	0.29
	(0.49)	(0.50)	(0.45)
Years of education	11.32	10.58	10.94
	(4.28)	(2.27)	(3.38)
Age at entry	26.76	28.59	29.79
	(8.72)	(8.86)	(9.42)
Years since migration	12.72	18.51	8.78
	(9.91)	(11.14)	(7.42)
Married	0.68	0.85	0.60
	(0.47)	(0.35)	(0.49)
One child	0.19	0.50	0.23
	(0.39)	(0.50)	(0.42)
Two children	0.19	0.21	0.22
	(0.39)	(0.41)	(0.42)
Three or more children	0.14	0.17	0.38
	(0.35)	(0.38)	(0.49)
Distance to home country (in km)	5759.67	1930.28	2147.70
	(3995.59)	(1527.57)	(2279.36)
Southern states	0.29		
	(0.45)		
Non-metropolitan area	0.01		
	(0.12)		
Minority language share	0.33		
	(0.25)		
Former colony	0.11		0.03
	(0.32)		(0.16)
Neighboring country		0.14	
		(0.35)	
Abroad five years ago	0.23		
	(0.42)		
Family abroad		0.30	
		(0.46)	
Refugee	0.12	0.06	
	(0.32)	(0.25)	
Political reasons			0.02
			(0.15)

Notes: – Number of observations: 59,890 in the 2000 U.S. Census, 698 in the SOEP, and 2,540 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	ASJP	ASJP	ASJP
	ME/StdE	ME/StdE	ME/StdE	ME/StdE
Linguistic distance (Test-score-based)	−0.003***			
	(0.000)			
Levenshtein distance (ASJP)		−0.002***	−0.004***	−0.002***
		(0.000)	(0.001)	(0.000)
Years of education	0.051***	0.050***	0.056***	0.033***
	(0.001)	(0.001)	(0.010)	(0.003)
Age at entry	−0.017***	−0.017***	−0.033*	−0.004
	(0.002)	(0.001)	(0.015)	(0.006)
Age at entry^2	0.000***	0.000***	0.000*	−0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Years since migration	0.017***	0.017***	0.022*	0.030***
	(0.001)	(0.001)	(0.010)	(0.004)
Years since migration^2	−0.000***	−0.000***	−0.001*	−0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Married	0.016**	0.016**	0.028	0.019
	(0.005)	(0.005)	(0.063)	(0.022)
Children in the HH. (Ref.= 0)				
One child	0.016*	0.018**	0.005	−0.068**
	(0.006)	(0.006)	(0.078)	(0.024)
Two children	0.014*	0.013*	−0.053	0.020
	(0.006)	(0.006)	(0.077)	(0.037)
Three or more children	0.001	0.003	−0.092	−0.107**
	(0.007)	(0.007)	(0.081)	(0.034)
Distance to home country (in km)	0.000***	0.000***	0.000*	−0.000*
	(0.000)	(0.000)	(0.000)	(0.000)
Distance to home country^2(in km)	0.000*	−0.000	−0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Southern states	0.059***	0.062***		
	(0.005)	(0.005)		
Non-metropolitan area	0.032†	0.025		
	(0.019)	(0.019)		
Minority language share	−0.438***	−0.493***		
	(0.016)	(0.017)		
Former colony	0.173***	0.158***		0.297***
	(0.007)	(0.007)		(0.082)
Neighboring country			0.331***	
			(0.058)	
Abroad five years ago	−0.107***	−0.107***		
	(0.007)	(0.007)		
Family abroad			−0.117*	
			(0.052)	
Refugee	−0.161***	−0.136***	−0.203*	
	(0.008)	(0.008)	(0.084)	
Political reasons				0.068
				(0.063)
Pseudo-R^2	0.245	0.240	0.132	0.138
Observations	59890	59890	698	2540

*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.*

Table 4: Descriptive Sample Statistics – International Trade Sample

	Mean	StdD	Min	Max
Log real trade	10.062	3.336	−16.09	20.81
Both in GATT/WTO	0.480	0.500	0.00	1.00
One in GATT/WTO	0.269	0.444	0.00	1.00
General system of preferences	0.231	0.422	0.00	1.00
Log distance	8.165	0.809	3.78	9.42
Log product real GDP	47.881	2.676	35.39	59.09
Log product real GDP p/c	16.034	1.504	9.72	21.60
Regional FTA	0.015	0.120	0.00	1.00
Currency union	0.014	0.118	0.00	1.00
Land border	0.031	0.172	0.00	1.00
Number landlocked	0.246	0.466	0.00	2.00
Number islands	0.341	0.540	0.00	2.00
Log product land area	24.206	3.280	9.64	32.77
Common colonizer	0.100	0.300	0.00	1.00
Currently colonized	0.002	0.044	0.00	1.00
Ever colony	0.021	0.142	0.00	1.00
Common country	0.000	0.017	0.00	1.00
Common language	0.223	0.416	0.00	1.00
Levenshtein distance	90.256	22.453	0.00	107.33
Levenshtein distance LF	75.063	36.787	0.00	105.81
Linguistic features index[a]	0.429	0.203	0.00	1.00

Notes: – Number of observations: 234,597 in 12,150 country-pair groups, except [a] 227,145 in 11,348 country-pair groups. – Observations per group within [1,52], mean = 19.3.

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

	ComLang Coef/StdE	LevDist I Coef/StdE	LevDist II Coef/StdE	LingFeat Coef/StdE
Both in GATT/WTO	−0.004	−0.003	−0.006	−0.018
	(0.034)	(0.034)	(0.034)	(0.034)
One in GATT/WTO	−0.135***	−0.134***	−0.133***	−0.154***
	(0.034)	(0.034)	(0.034)	(0.034)
General system of preferences	0.697***	0.720***	0.699***	0.709***
	(0.032)	(0.031)	(0.031)	(0.032)
Log distance	−1.308***	−1.274***	−1.303***	−1.288***
	(0.023)	(0.024)	(0.023)	(0.024)
Log product real GDP	0.188***	0.186***	0.185***	0.182***
	(0.052)	(0.052)	(0.052)	(0.053)
Log product real GDP p/c	0.521***	0.522***	0.523***	0.539***
	(0.049)	(0.049)	(0.049)	(0.050)
Regional FTA	0.935***	0.936***	0.933***	0.969***
	(0.127)	(0.126)	(0.126)	(0.130)
Currency union	1.199***	1.276***	1.195***	1.231***
	(0.122)	(0.125)	(0.123)	(0.124)
Land border	0.277*	0.280**	0.281**	0.290*
	(0.108)	(0.108)	(0.108)	(0.113)
Number landlocked	−1.002***	−1.109***	−1.111***	−0.934***
	(0.254)	(0.252)	(0.251)	(0.258)
Number islands	−0.828***	−0.927***	−0.877***	−0.881***
	(0.189)	(0.190)	(0.192)	(0.198)
Log product land area	0.372***	0.366***	0.361***	0.366***
	(0.033)	(0.032)	(0.033)	(0.033)
Common colonizer	0.607***	0.702***	0.595***	0.687***
	(0.064)	(0.062)	(0.065)	(0.065)
Currently colonized	0.777**	0.779**	0.787**	0.748**
	(0.263)	(0.253)	(0.264)	(0.262)
Ever colony	1.268***	1.266***	1.256***	1.330***
	(0.114)	(0.116)	(0.114)	(0.113)
Common country	0.259	0.066	0.235	0.255
	(0.580)	(0.653)	(0.576)	(0.613)
Common language	0.265***			
	(0.043)			
Levenshtein distance		−0.006***		
		(0.001)		
Levenshtein distance LF			−0.003***	
			(0.000)	
Linguistic features index				−0.606***
				(0.098)
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	1.819	1.818	1.819	1.807
F Statistic	274.05***	271.77***	274.59***	265.43***
Observations	234597	234597	234597	227145

*Notes: – Significant at: ***0.1% level; **1% level; *5% level; †10% level. – Robust standard errors (clustering by country-pairs) are reported in parentheses. – 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).

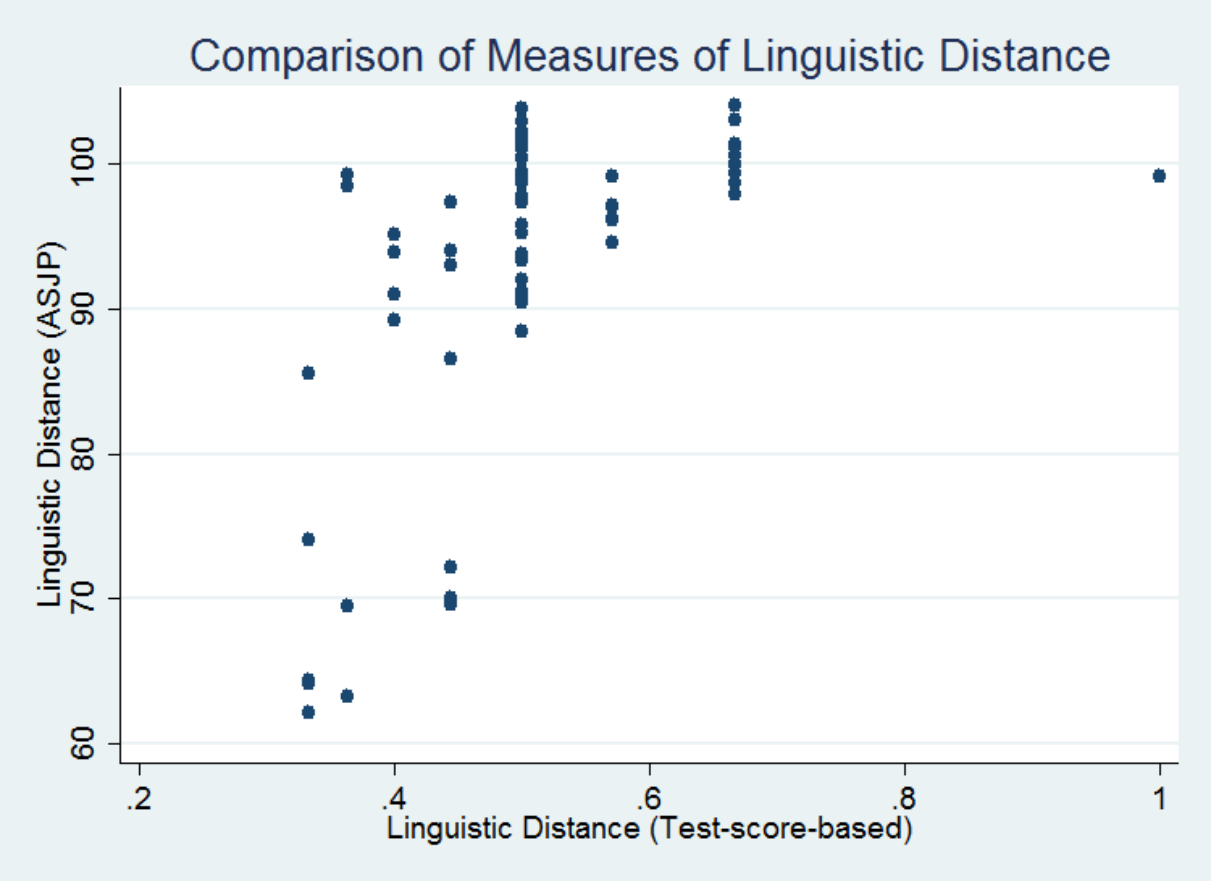

Figure A1: Comparisons of Linguistic Distance Using Test-Score-Based and ASJP Approach – Immigration Sample

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.

Sensitivity Analyzes – International Trade Sample

Tables A3 and A4 present some of the sensitivity analysis we have performed. They confirm that our key results do not depend delicately on the sample used in the estimation. Table A3 examines the sensitivity of the results with respect 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 A4 analyzes the sensitivity of our results when we restrict our sample to the slightly smaller one of Lohmann's linguistic features index.

Table A3: Effect of Language on Bilateral Trade – OLS Results, Subsample Language Barrier > 0

	LevDist I Coef/StdE	LevDist II Coef/StdE	LingFeat Coef/StdE
Both in GATT/WTO	−0.006	−0.038	−0.011
	(0.035)	(0.036)	(0.035)
One in GATT/WTO	−0.126***	−0.139***	−0.147***
	(0.035)	(0.036)	(0.036)
General system of preferences	0.744***	0.601***	0.673***
	(0.031)	(0.032)	(0.031)
Log distance	−1.278***	−1.208***	−1.268***
	(0.025)	(0.027)	(0.027)
Log product real GDP	0.072	−0.054	0.011
	(0.053)	(0.058)	(0.057)
Log product real GDP p/c	0.644***	0.786***	0.721***
	(0.051)	(0.056)	(0.054)
Regional FTA	0.823***	−0.276*	0.172
	(0.148)	(0.119)	(0.156)
Currency union	1.329***	1.165***	1.223***
	(0.133)	(0.274)	(0.203)
Land border	0.280*	0.388**	0.300*
	(0.120)	(0.130)	(0.133)
Number landlocked	−0.064	0.029	−1.509***
	(0.314)	(0.252)	(0.263)
Number islands	0.064	−0.440*	−2.101***
	(0.193)	(0.211)	(0.224)
Log product land area	0.535***	0.600***	0.579***
	(0.039)	(0.042)	(0.040)
Common colonizer	0.696***	0.884***	0.661***
	(0.063)	(0.092)	(0.069)
Currently colonized	0.348	1.183**	0.460
	(0.292)	(0.388)	(0.387)
Ever colony	1.514***	1.040***	1.180***
	(0.131)	(0.193)	(0.152)
Common country	1.181***		
	(0.346)		
Levenshtein distance	−0.008***		
	(0.002)		
Levenshtein distance LF		−0.008***	
		(0.002)	
Linguistic features index			−0.262*
			(0.123)
Year fixed effects	yes	yes	yes
Country fixed effects	yes	yes	yes
Adjusted R^2	0.702	0.706	0.707
RMSE	1.828	1.793	1.802
F Statistic	856.82***	267.18***	275.74***
Observations	223580	191368	205756

*Notes: – Significant at: *** 0.1% level; ** 1% level; * 5% level; † 10% level. – Robust standard errors (clustering by country-pairs) are reported in parentheses. – 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 of the equations because of collinearity.*

Table A4: Effect of Language on Bilateral Trade – OLS Results, Subsample Linguistic Features Index

	ComLang Coef/StdE	LevDist I Coef/StdE	LevDist II Coef/StdE
Both in GATT/WTO	−0.016	−0.014	−0.017
	(0.034)	(0.034)	(0.034)
One in GATT/WTO	−0.151***	−0.149***	−0.147***
	(0.034)	(0.034)	(0.034)
General system of preferences	0.695***	0.719***	0.696***
	(0.032)	(0.031)	(0.032)
Log distance	−1.300***	−1.263***	−1.292***
	(0.023)	(0.024)	(0.023)
Log product real GDP	0.189***	0.187***	0.187***
	(0.053)	(0.053)	(0.053)
Log product real GDP p/c	0.533***	0.533***	0.535***
	(0.050)	(0.050)	(0.050)
Regional FTA	0.973***	0.974***	0.969***
	(0.129)	(0.129)	(0.129)
Currency union	1.210***	1.290***	1.197***
	(0.123)	(0.126)	(0.123)
Land border	0.284*	0.289**	0.291**
	(0.113)	(0.112)	(0.112)
Number landlocked	−0.829**	−0.946***	−0.947***
	(0.264)	(0.261)	(0.259)
Number islands	−0.828***	−0.938***	−0.891***
	(0.196)	(0.197)	(0.199)
Log product land area	0.373***	0.367***	0.360***
	(0.033)	(0.033)	(0.033)
Common colonizer	0.597***	0.701***	0.572***
	(0.067)	(0.065)	(0.068)
Currently colonized	0.763**	0.764**	0.773**
	(0.267)	(0.256)	(0.269)
Ever colony	1.249***	1.244***	1.216***
	(0.115)	(0.117)	(0.114)
Common country	0.282	0.082	0.257
	(0.592)	(0.672)	(0.589)
Common language	0.285***		
	(0.044)		
Levenshtein distance		−0.006***	
		(0.001)	
Levenshtein distance LF			−0.004***
			(0.001)
Year fixed effects	yes	yes	yes
Country fixed effects	yes	yes	yes
Adjusted R^2	0.705	0.705	0.705
RMSE	1.807	1.806	1.806
F Statistic	268.12***	265.47***	268.95***
Observations	227145	227145	227145

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