Speaking of Trade: Quantifying the Contribution of Multilingualism to Overcome the Language Barriers to Trade[1]

Adrienne Mack
Mutual of Omaha

Enrique Martínez-García
Federal Reserve Bank of Dallas

María Teresa Martínez-García
University of Kansas

December 2014

Abstract

There is growing recognition that language affects bilateral trade through multiple channels, as most gravity trade models include a language variable. Previously, the focus has been on common (official or spoken) native language or more gradient measures of linguistic proximity across native languages, all of them showing statistically positive effects on bilateral trade. This paper explores the impact of nonnative languages on trade. We find that the effect of indirect communication through a nonnative language is larger than that of a shared native spoken language. These results suggest evidence of the emergence of regional/global trading languages (*lingua franca*), which may help reduce the language barriers to trade.

JEL codes: F10, F40
Keywords: Language; Lingua Franca; International Trade; Gravity Model of Trade

[1] Enrique Martínez-García (corresponding author), Federal Reserve Bank of Dallas, 2200 N. Pearl Street, Dallas, TX 75201. 214-922-5262. enrique.martinez-garcia@dal.frb.org. https://sites.google.com/site/emg07uw/. Adrienne Mack, Mutual of Omaha. adrienne.mack@ymail.com. María Teresa Martínez-García, University of Kansas, Department of Linguistics, 1541 Lilac Lane, Blake Hall, Room 427, Lawrence, KS 66045-3129. 785-864-3450. maria.martinezgarcia@ku.edu. We thank Jacques Melitz and Farid Toubal for facilitating their data as a reference and for their pioneering work on linguistic data. We would also like to thank J.B. Cooke and Christoffer Koch, and participants in the Spring 2014 Midwest Macro Meeting and the 89th Annual Western Economic Association International (WEAI) conference for helpful suggestions. We gratefully acknowledge the research assistance provided by Valerie Grossman, the help of Bradley Graves, and the support of the Federal Reserve Bank of Dallas. This research was initiated while María Teresa was visiting and Adrienne Mack was working at the Federal Reserve Bank of Dallas, whose hospitality she greatly appreciates. All remaining errors are ours alone. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Dallas or the Federal Reserve System.

1 Introduction

Language barriers consistently appear in gravity models as a significant deterrent to trade for a variety of reasons—among them are the costs associated with communicating in a foreign language. Sharing an official or native language is generally found to facilitate trade (see, e.g., Egger and Lassman (2012)). As documented by Lohmann (2011) and Melitz and Toubal (2014), speaking similar native languages also reduces language barriers since it becomes easier to learn the other country's language.

Controlling for economic size, trade is inversely proportional to distance according to the basic gravity model—as illustrated in Figure 1. However, we are starting to recognize in the data that, apart from sharing a native language, being able to communicate using the other country's native language or a third language commonly-spoken can also facilitate trade. Being able to communicate in a second language, not only allows for the possibility of increasing the number of trading partners with those that speak, as native speakers, the language that two countries share, but also allows for the possibility of trading with third countries, which have also a significant share of the population able to speak it as a second language.

We can show also the role of multilingualism with the three examples illustrated in Figure 1: Spain, Brazil, and Japan. Using only direct communication, Spain appears to trade with a total of 21 countries that share Spanish as their native (or official) language to a larger extent that would be predicted by the basic gravity model. This is an example of the standard result documented in the literature where it is found that sharing a common native language can facilitate trade between two partners.

Trading with Spanish-speaking countries supposes a reduction in costs (e.g., translation costs), as it implies having a relatively big market share without the need of intermediaries (e.g., translator/interpreters). However, Spanish companies can make the decision of whether investing resources in translation or acquiring skills in other languages to increase their trading options beyond countries that share the Spanish language as their native language. If Spanish companies are able to communicate in English, this could imply that they have access to English-speaking countries (which adds 67 countries), but also to third countries that also have knowledge of English.

Brazil shows a slightly different story. Portuguese is only spoken as a native (or official) language in 7 countries, most of which are not located within close distance of Brazil itself. Brazil has the possibility to trade with these other Portuguese-speaking countries. However, this is indeed a rather small market for the size of the Brazilian economy. It is, then, reasonable to assume that investing in some sort of indirect communication can be beneficial for the country's economy. Not surprisingly, what we observe is that common spoken languages (other than Portuguese, although primarily we refer to Spanish) appear the same role in facilitating trade than the native language did for Spain.

In the case of Japan, it is particularly useful to improve trading opportunities by increasing the ability to communicate in a foreign language, given that Japanese is not spoken in other countries. For Japan, being able to communicate in a second language (English in this case), not

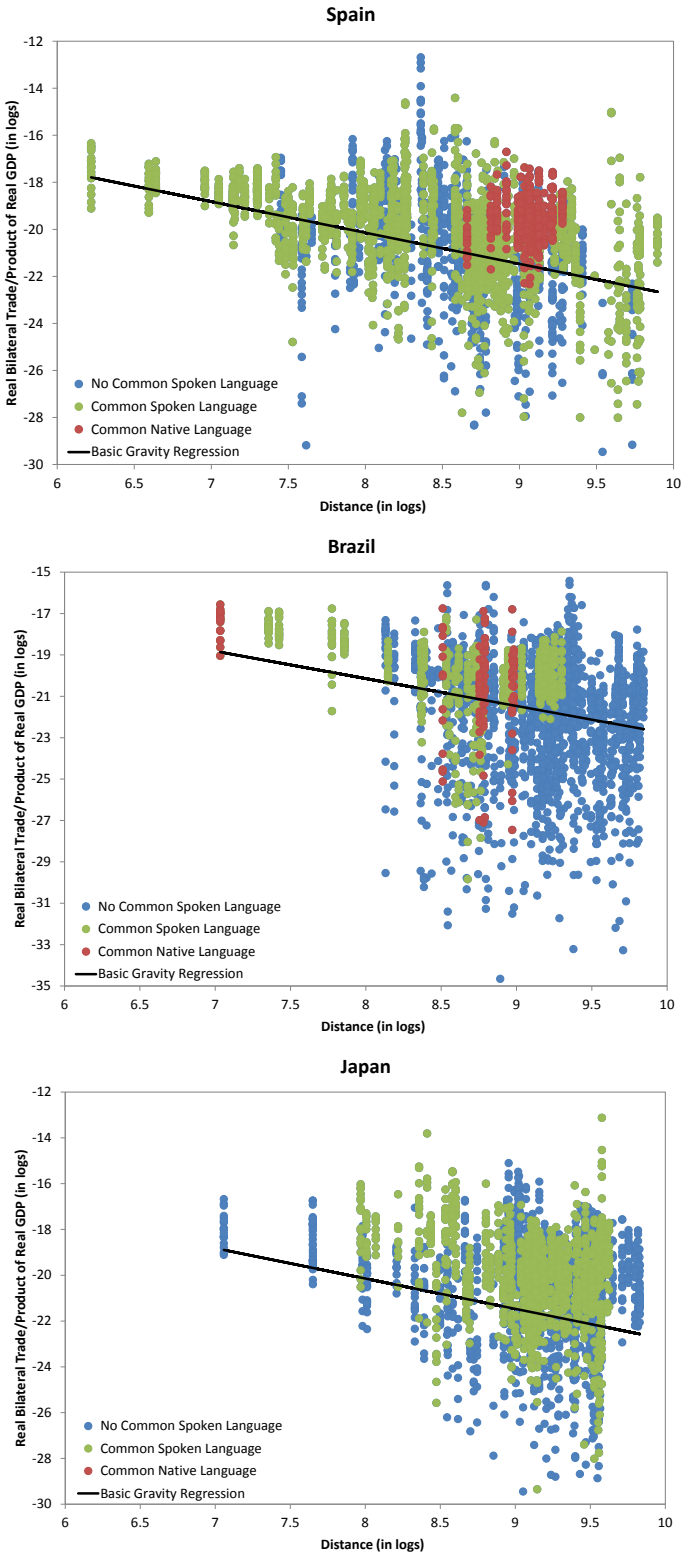
only allows them the possibility of more easily reaching foreign markets with their products (e.g., the ability of Japanese businesses to trade with American companies using English), but also allows for the possibility of trading with third countries, which have also adopted English as a second language (e.g., the possibility of Japanese and Chinese companies to trade using English).

Similarly, half of the population in Papua New Guinea (a former British colony) speaks English as a second language, but less than 4 percent are native-speakers of English. Looking at native languages alone would not fully capture Papua New Guinea's ability to trade with other English-speaking (native or otherwise) trading partners, and therefore would not provide us with a complete picture on how languages affect trading relationships.

In this paper, we use data on spoken languages to investigate how the ability to speak non-native language(s) impacts bilateral trade. While the trade effect of sharing a common (spoken) native language remains statistically significant and sizeable, we show that the impact of indirect communication through a language that is non-native for one of the trading partners can be even larger. We interpret this as evidence that the emergence of regional trading languages (see, e.g., Melitz (2008)) and possibly a world *lingua franca* in the case of the English language (see, e.g., Crystal (2003) and Ginsburgh and Weber (2011)) has greatly reduced the language barriers to trade.

The paper is organized as follows: In Section 2 we describe the gravity model and its basic implications. Section 3 describes the econometric methodology, while section 4 reviews the linguistic data available. Section 5 discusses our empirical findings on the implications of language for trade, while Section 6 concludes.

Figure1. Illustration of the Language Factor in the Basic Gravity Model of Trade



Note: The plotted gravity regression line is estimated with the full sample (154042 observations): constant = -9.556 (s.e. 0.067), slope on log-distance = -1.324 (s.e. 0.008), Adjusted R^2 = 0.140.

2 The Gravity Model of Trade

The basic gravity model of trade dating back to Tinbergen (1962) and Pöyhönen (1963) explains trade flows between two countries as directly proportional to their respective economic sizes (measured by their GDP) and inversely proportional to their geographic distance. The derivation of the gravity model in general equilibrium can arise from a potentially large class of theoretical models (see, e.g., Deardorff (1998) and Head and Mayer (2015)). The general form of the export demand equation obtained in general equilibrium can be written as follows,

$$x_{ij} = \frac{y_i y_j}{y^W}\left(\frac{\tau_{ij}}{\Lambda_i P_j}\right)^{1-\sigma}, \tag{1}$$

where y_j is the nominal income in country *j*, $y^W = \sum_j y_j$ is the world nominal income, $\phi_j = \frac{y_j}{y^W}$. defines the income share of country *j*, and x_{ij} is the nominal value of exports. Moreover, the term Λ_i can be expressed as,

$$\Lambda_i = \left(\sum_j \left(\frac{\tau_{ij}}{P_j}\right)^{1-\sigma} \phi_j\right)^{\frac{1}{1-\sigma}}, \tag{2}$$

and the consumer price index as,

$$P_j = \left[\sum_i \left(\frac{\tau_{ij}}{\Lambda_i}\right)^{1-\sigma} \phi_i\right]^{\frac{1}{1-\sigma}}. \tag{3}$$

Taken together, equations (2) and (3) form a system that expresses Λ_i and $P_i \ \forall i$ in terms of nominal country income shares $\{\phi_i\}$, bilateral trade costs $\{\tau_{ij}\}$, and the elasticity of substitution σ.

The export equation in (1) subject to the price indexes given in (2)-(3) (also referred as the multilateral resistance terms) represents the general equilibrium form of the gravity model. The gravity model shows that bilateral trade, after controlling for the economic size of the trading partners (given by their nominal incomes), depends on the bilateral trade costs τ_{ij} relative to their multilateral resistance terms P_i and P_j. The vector of trade costs $\{\tau_{ij}\}$ in (1) and (2)-(3) is considered to include transportation costs and tariffs as well as other costs that are harder to observe and measure such as information and networking costs, costs of design adaptations for the export market, legal and regulatory costs, cultural, institutional and linguistic barriers, etc. Prices differ across locations due to these largely unobserved trade costs, so their identification

constitutes one of the key aims of empirical trade work that we tackle in this paper. One must also confront the fact that the multilateral resistance terms $\{P_i\}$ and $\{P_j\}$ are also unobserved.

The "canonical" version of the gravity model employed in the empirical literature does not include the multilateral resistance terms, only distance (as a proxy for the unobserved trade costs) and the two countries' nominal incomes (measured by their respective GDPs). The advantage of the general equilibrium derivation in (1) and (2)-(3) is that it shows how accounting for the role of bilateral trade costs while omitting the impact of the multilateral resistance terms can bias the inferences we derive. The multilateral resistance terms are themselves endogenously determined and depend on the bilateral trade costs with all countries simultaneously.

There are three main approaches in the literature to handle the unobserved nature of $\{P_i\}$ and $\{P_j\}$. One possibility would be to use measured price indexes as proxies for the multilateral resistance terms, but the mapping between data and theory remains a concern and a source of measurement error. Another possibility would be to estimate (1) and (2)-(3) jointly as Anderson and van Wincoop (2003) among others have tried to do, although that requires the use of nonlinear methods. Finally, one can also use fixed effects to proxy for these multilateral resistance terms as well as other location-specific factors that do not vary across trading countries in the tradition of Rose and van Wincoop (2001) or Eaton and Kortum (2002). For tractability and because of misspecification concerns, our own estimation strategy modifies the "canonical" version of the gravity model based on the fixed-effect approach to control for the multilateral resistance terms.

Using solely distance as a proxy for the unobservable trade costs appears to capture only part of what these trade costs represents. We assume that bilateral trade costs $\{\tau_{ij}\}$ take the following parametric functional form that includes other observables to proxy other than distance as proxies,

$$\tau_{ij} = g\left(D_{ij}, \{Z_{ij}^{\ m}\}, \{Q_{ij}^{\ k}\}\right) = \left(D_{ij}\right)^{\gamma_0} \prod_m \left(Z_{ij}^{\ m}\right)^{\gamma_m} \exp\left(\sum_k \nu_k Q_{ij}^{\ k}\right), \tag{4}$$

where D_{ij} is the bilateral distance between country *i* and country *j*, the vector $\{Z_{ij}^{\ m}\}$ contains other variables that help us proxy for bilateral trade costs except the dummy variables that are collected in the vector $\{Q_{ij}^{\ k}\}$. In this case, γ_0 and $\{\gamma_m\}$ represent their corresponding elasticities while $\{\nu_k\}$ is a vector of semi-elasticities.

The "remoteness" variables such as distance are important because more remote countries have to offer a lower price to their trading partner in order to compensate for higher transportation costs, and this has been broadly understood in the empirical literature as distance figures prominently in the "canonical" model. Apart from distance, we also consider a variable that identifies whether the trading partners share a common land border in order to more precisely pin down the border effect in our estimates. While the relative size of the countries is pinned down in

the gravity model with the nominal income's of the trading partners, we also consider the log of product of land area as a further control

The general equilibrium version of the gravity model presented here in equations (1) and (2)-(3) offers further insight on bilateral trade costs in general and distance in particular. It can be shown, as noted by Anderson and van Wincoop (2003), that bilateral trade x_{ij} between countries *i* and *j* is homogeneous of degree zero in the bilateral trade costs τ_{ij}, and the equilibrium multilateral resistance terms are themselves homogenous of degree ½ in the trade costs. The key implication is that *relative* trade barriers between countries (that is, the bilateral costs relative to the average trade costs that each country faces with all its trading partners) is what determines trade volumes rather than the absolute level of the trade barriers themselves. If one conceives that the decline in transportation and shipping costs over the past decades extensively discussed in the literature has constituted a rather uniform phenomenon, this homogeneity result may explain why standard estimates of the gravity model have not detected a decline in the importance of distance (as the closest proxy variable for transportation costs) over time.

We focus our attention, in turn, on other proxy variables $\left\{Z_{ij}^{\ m}, Q_{ij}^{\ k}\right\}$ that describe other costs affecting the bilateral trade through τ_{ij}. For instance, we control for standard legal, institutional and regulatory costs with variables that identify participation in a currency union and/or currency board, a common free trade area (FTA), political union or former colonial relationship, and the ranking of economic freedom between the trading partners. However, our primary interest is in understanding the existing measures of measures of linguistic "remoteness" (or similarity) instead. Linguistic distance between trading partners varies rather slowly over time, but we otherwise their effect could be compounded and confounded with the border effect picked up our purely geographical distance variables. The case for linguistic distance is often argued in the literature (see, e.g., Helliwell (1999), Anderson and van Wincoop (2004) and Melitz (2008) among others), but in the next section we shall propose a stylized model of language to ground on theory our case for including these variables. To the best of our knowledge, this is one of the first attempts to model language explicitly in an economic environment (which we think constitutes a valuable contribution to the literature by itself).

We estimate the gravity model in logs, but in the literature has also investigated the model in levels. See Santos Silva and Tenreyro (2006) for a further discussion of both methods. After replacing (4) into (1) for the unobserved bilateral trade costs and introducing fixed effects to control for the multilateral resistance terms, we represent the gravity equation in log-linear form as follows,

$$\ln\left(x_{ij}\right) = \ln\left(k\right) + \ln\left(y_i\right) + \ln\left(y_j\right) + \left(1-\sigma\right)\left(\gamma_0 \ln\left(D_{ij}\right) + \sum_m \gamma_m \ln\left(Z_{ij}^{\ m}\right) + \sum_k \nu_k Q_{ij}^{\ k} - \alpha_i^P - \alpha_j^P\right) + \eta_{ij}, \quad (5)$$

where the constant *k* captures $k = \frac{1}{y^w}$, while the fixed effect terms α_i^p and α_j^p incorporate unobserved location-specific attributes that affect the multilateral resistance terms. Equation (5)

is the basis for all the estimations we present below. Our goal is to assess the relevance of the border effect when we account for linguistic remoteness in the gravity model, but as the log-linear version of the model makes clear we should note that the estimates we derive would be a reduced-form that combines the elasticities of $\{\gamma_0, \gamma_m\}$/semi-elasticities of $\{\nu_k\}$ with the preference parameter σ that determines the elasticity of substitution across varieties.

3 Methodology

The basic gravity model relates log-real bilateral trade x_{ijt} to the log-product of the real GDP of each country pair $\left(y_{it}y_{jt}\right)$ and the log of their distance D_{ijt} (see Tinbergen (1962)). Distance serves as a proxy for transportation cost. As the distance between two bilateral trading partners increases, the cost of trade increases. If the cost becomes too large and the finical burden is too great, trade is no longer an option. This causes the average trade volume between two countries to decline. GDP is included to address the fact that wealthier nations are better able to withstand a cost increase, given the higher demand. However, distance and GDP alone cannot explain all the reasons that one country would choose to trade with another.

There are many other factors that go into selecting a trading partner, such as geographical location, ease of travel, historical ties, a familiar ideological system, etc. These must also be controlled for in order to isolate the role of language and correctly measure its effect on trade. Once isolated, various measures of language can be introduced into the equation to identify the specific roles language plays in influencing bilateral trade decisions. We add two additional variables to control for other important channels in which language can affect trade, literacy and linguistic diversity. Often trade agreements must be established through written contracts. The ability to verbally communicate to another country does not guarantee the ability to conduct business through written communication. To control for non-verbal communication, we include in our gravity regression a measure of literacy. We also add measure of linguistic diversity. This measures the number of different languages currently spoken in a country, since the trade effect of a country's ability to learn an additional language might vary depending on how many languages are already spoken.

In the augmented specification in (1), we add language variables ${L_{ijt}}^m$ and other geographic, historic (colonial) and economic variables ${Q_{ijt}}^k$ to the basic gravity equation.

$$\ln\left(x_{ijt}\right) = \beta_0 + \beta_1 \ln\left(y_{it}y_{jt}\right) + \beta_2 \ln\left(D_{ijt}\right) + \sum_m \gamma_m {L_{ijt}}^m + \sum_k \nu_k {Q_{ijt}}^k + \eta_{ijt}. \quad (6)$$

We estimate the model in logs with panel data for 196 countries from 1970 to 2010 in five year intervals, including time country-specific and time fixed effects. We apply the iterative algorithm implemented in Stata® by Guimaraes and Portugal (2010) to solve the two-way Fixed Effects (FE) problem with unbalanced data and very large numbers of effects through the **reg2hdfe** command which also allows for clustered standard errors.

All bilateral trade data comes from the U.N. COMTRADE database, is quoted in U.S. dollars and deflated with the U.S. CPI. The real GDP (and real GDP per capita) are obtained from the Penn World Tables database complemented with other sources. The geographic variables are constructed using national sources, and distance between each country's economic center is calculated with the Great-circle formula (in kilometers).

All linguistic variables are constructed from primary sources such as Ethnologue (see, e.g., Lewis (2009)) and the CIA (2011), and completed with national sources. All other variables are extended using original data from the Centre d'Études Prospectives et d'Informations Internationales (CEPII). See the appendix for more information.

4 Linguistic Data

As in Melitz (2008) and Melitz and Toubal (2014), we consider only 42 languages spoken by at least 4 percent of the population in at least 2 countries (see Table 1 in Melitz and Toubal (2014)). We construct the common spoken language index (CSL) with all non-native (e.g., immigrant languages and second-languages) and native languages spoken in each country pair. Having obtained the population shares of speakers per country from Melitz and Toubal (2014) and Ethnologue (see, e.g., Lewis (2009)), we construct a Boisso and Ferrantino (1997)-type indicator of linguistic distance for each country pair as,

$$CSL_{ij} \equiv \sum_{n=1}^{42} s_{in} s_{jn}, \tag{6}$$

where s_{in} and s_{jn} are the population shares that speak language n (as a first- or second-language) in country i and country j respectively.

A higher CSL index means either a higher probability that two individuals from both countries in a given pair are able to communicate in a particular shared language, or a higher number of common spoken languages among the county pair. The CSL is additive in the product of the shares, so we can decompose the measure into three categories: spoken languages that are native in both trading partners, spoken languages that are native in one country but not the other, and spoken languages that are non-native in both. This decomposition of CSL allows us to disentangle the effect of learning the language of your trading partners or a third language to facilitate trade.

Native languages are derived as a subset of the spoken languages, constructed from the same sources but only including those spoken by at least 4 percent of the population as their first-language.[2] We construct common native language (CNL) as a binary variable that takes the value of 1 whenever the trading partners share at least one native language and 0 otherwise. The common official language (COL) variable is constructed to allow at most 2 official languages per country, using primarily CIA (2011) data. COL assigns the value 1 if two countries share a common official language and 0 otherwise.

[2] Native language is derived from the number of first-language speakers of that given language in the country. It allows for the possibility of bilingual speakers, in which case they would be counted for each language they speak.

We investigate variants linguistic proximity on a lexical scale with a variant of the two measures proposed by Melitz and Toubal (2014). Our LP1 and LP2 measures assign the value of 1 whenever at least 20 percent of the populations share the same native language. For all other country pairs, LP1 calculates linguistic proximity on the basis of the Ethnologue (Lewis (2009)) language trees,[3] and LP2 uses the statistical analysis of lexical similarity adjusted for noise from the Automated Similarity Judgment Program (ASJP) project (Holman (2011)).

5 Empirical Findings

We run the regression imposing unit elasticity to be consistent with the theoretical framework, and we use fixed effects, as they provide unbiased estimates of the partial effect coefficients arising from the bilateral pair variables. These fixed effects should subsume all the country characteristics: in this regards, log-product of population and/or GDP per capita, the log-product of the areas, and the landlocked variables were all considered in the analysis but excluded. We also consider two important country characteristics with connection to the linguistic effect on trade (literacy and diversity) that are not fully captured by the fixed effect specification and augment the model for the sub-sample after 1995.

Dummies for the common official, common native and common spoken (native, native-nonnative, nonnative) variables were used because they provide a better sense of whether a given language was spoken over the entire period or not. However, we could not compute reliable indicators of the spoken language based on the share of the population that speaks the language or include other important aspects of the language such as similarity, literacy or diversity. In other words, we can safely assume that there is no time-variation within the sample period considered in this paper for whether two countries share a common official/native/spoken language or not; however, there are reasons to believe that over time the extent to which a given second language is spoken in a particular country may change. We also explore these indicators on a shorter sample (starting in 1995) for which the assumption of no time-variation in the variables appears more plausible.

All Tables below contain the main tests of our gravity equation with fixed effects. The estimated coefficients for the time- and country-fixed effects are omitted. The Student t's are based on robust standard errors (after correction for clustering of data for individual trading pairs). The estimated gravity model includes a number of variables that control for geographic features, existing economic relationships, cultural relationships and historical ties between two trading partners. Omitting them introduces a clear upward bias on our estimates of language, since language is often thought to capture the effect of pre-established cultural and historical ties. All coefficients tend to be strongly statistically significant, generally unchanged across different specifications, and of the expected sign. We interpret this as indicating that cultural, political and historical ties do not fully account for the role of language in trade.

[3] We assign 0 for 2 languages belonging to separate family trees, 0.25 for 2 languages in different branches of the same family tree (Hindi and Greek), 0.50 for 2 languages in the same branch (Hindi and Farsi), 0.75 for 2 languages in the same sub-branch (Bengali and Hindi), and 1 for those countries sharing the same native language.

5.1 Official, Native, and Spoken Language

Table 1 tries to determine whether there is a partial effect that can be attributed to speaking a foreign language and whether this effect is robust across time or not. From this exercise, we can point out that common official language alone combines two effects—the fact that a language is official and the fact that most official languages are either native or at least spoken in the country. An official language is a language that is given a special legal status in a particular country. It is the language used to conduct government operations, but it does not typically refer to the language used in everyday communication. The results for that case are reported in models (1)-(3) that are estimated for the full sample and the pre-1990 (1970-1990) and post-1990 (1995-2010) periods, respectively. Model (1) and Model (2) show that sharing a common native language has a positive effect on trade, but that the effect of sharing an official language tends to dominate that of sharing a common native language. This suggests that communication in a language with official status recognition is favored over communication in a non-official native language. When we incorporate either a common native language dummy (Models (4)-(6)) or a common spoken language dummy (Models (7)-(9)) we find that the partial effect of official status stands around 0.3 and is remarkably stable across time, but also robust to whether we look at native languages alone or at the broader set of all spoken languages. The common spoken language variable introduced in Model (10), Model (11), and Model (12) presents a broader measure of linguistic distance that includes all major languages spoken in each trading partner. In addition to native languages, it includes immigrant or second languages, thus capturing non-native languages that can be used as trading languages.

Contrary to the stability found with the coefficient on the official status of the language, we observe a significant change over time in the estimated partial effect from the native and spoken language variables. The results we estimate for the period after 1995 conform to those reported for total trade and a similar sample period by Melitz and Toubal (2014). We find, however, that the estimated coefficients on native or spoken language have become larger and more statistically significant in the post-1990 period while the partial effects on official status have remained largely unchanged given that we explore a panel with a longer time series coverage than that investigated in Melitz and Toubal (2014). In other words, sharing a common native or spoken language has become much more important since the 1990s. The mechanism through which language affects trade has, indeed, evolved over time. As trade expanded during this period, the evidence on partial effects suggests that it has increasingly been directed towards trading partners with whom we share a common language. One possible explanation for this phenomenon could be that the impact of language might be more relevant for the extensive than the intensive margin. Although language barriers can add to the fixed costs of entry into a given market, they also must have an effect on the variable costs of operating into that market once the decision to export has been made. In this sense, language barriers may have changed over time but it might also be the case that as overall trade increases the composition is shifted towards countries with whom there are lower linguistic trade barriers.

The more interesting results appear in Models (10)-(12). All native languages of a country are spoken languages too, so the common native language dummy is a subset of the indicator of common spoken native language. We can disentangle the common spoken language indicator naturally into three components. The first one captures the effect of sharing a common spoken

language that is native to both countries. The common spoken language dummy for native languages is indeed akin to the common native language dummy that we have included in Models (4)-(6). This component captures the possibility of direct communication between the populations of both trading partners on a language that is native to both.

The second and third components in which we split the information underlying the common spoken language are the novel contributions of our investigation and our primary focus since they allow us to test the partial effect that communication through a non-native language has on trade, beyond the role of sharing a common native/official language. To our knowledge, this is the first paper that has evaluated specifically the impact on trade of communication through a second language that is non-native to the country. The second component indicates instances where communication on a common spoken language occurs involving a language that is native to one partner, but nonnative to another. For instance, this dummy variable would take the value of one because for the bilateral pair between Papua New Guinea and Australia given that English is spoken (but nonnative) to Papua New Guinea and is spoken (and native) to Australia. The third dummy that we construct assigns the value of one to bilateral pairs for which there is the possibility of communicating through a language that is nonnative to both, but spoken by a significant share of the population that reaches at least 4 percent.

Our findings reported through Models (10)-(12) provide concrete evidence that communication through a nonnative second language has a statistically significant effect separate from whether a native or official language is shared between the two trading partners. The most interesting aspect about the results reported in the paper is that not only the effect of a common native language has been accentuated in the period post-1990, but also the impact of a second language has changed. In the period prior to 1990, the effect of a native language is small and not statistically significant. Sharing a common spoken language that is nonnative to at least one of the trading partners appears to impede trade to some extent during this period, and the effect is statistically significant in the case in which the common language is native to one of the trading partners. The only partial effect that matters and contributes to facilitate trade is that of sharing a common official language. One interpretation of the findings is that, during the 1970s and 1980s, education and training on a second language was actually determined by geopolitical considerations in a number of these countries. As a case in point, the citizens of the former Soviet Union would learn English, but the goal of English education was other than to facilitate trade with English-speaking countries of the West. In this regard, a common official language would be a better predictor of trade between countries during this period than speaking a second language. Over time, the realignment of the world that took place during the late 1980s and early 1990s lead many countries around the world to favor education on foreign languages and to allow more flexibility in the choice of foreign languages taught. This would, in turn, allow for the possibility of progressively shifting the education efforts towards languages that are most widely spoken and that have become historically major trading languages either globally (primarily English) or at regional levels (Spanish, French, Portuguese, Arabic, or Russian).

5.2 The Role of Cultural, Historical and Economic Ties on Trade

Related to the role of non-language variables, Table 1 shows that the partial effects estimated for geodesic distance and common border are relatively stable over time, as expected.

Among the variables included to capture the extent of colonial effects in the data (comcol, col45, colony, curcol), the most notable change occurs with the variable curcol that becomes quite large and weakly statistically significant in the post-1990 period. The estimated coefficient is negative, possibly indicating that countries currently involved in a colonial relationship tend to face additional hurdles to trade. Only variables that reflect historical colonial ties tend to facilitate trade nowadays as these historical ties tend to lead to stronger cultural and even linguistic ties in the present that can contribute to enter into a market and export easier for the firms of both countries. Sharing a common legal origin or greater religious proximity both tend to facilitate trade, and those findings have remained remarkably robust over time. Similarly, years at war has had a seemingly stable contribution to trade although in this case it tends to make it more difficult.

Among the most noticeable changes over time, as shown in Table 1, we can point out to the effect of regional trade agreements and the effect of sharing a common currency. Both show the most significant reversals in the post-1990 period. Up until 1990, results suggest that the effect of a regional trade agreement was negligible, while sharing a common currency had a very significant impact (as noted in the work of Rose (2002)). Presumably the trade agreements in place during the 1970s and 1980s were motivated by other geopolitical considerations and their efficacy to promote trade was hampered by that. In turn, sharing a common currency would appear to have a very large and significant effect. A lot more regional and targeted trade agreements were introduced later on, so trade was directed by means of these agreements which could explain why they became more important in our sample. In turn, the ex post effectiveness of sharing a common currency has significantly decline to the point of becoming negligible. For example, the first countries that decided to be involved in sharing a common currency would involve countries that are strongly related to each other through solid economic linkages. The major currency union of the post-1990 period was the euro, involving most of the member states of the European Union. In this case, having a common currency does have a political dimension too and the efficacy of it to promote trade within those countries appears to have been somewhat limited according to our data. However, once we control for the effect of other variables, including country specific characteristics, we do not find a reliable effect on trade from the currency union dummy.[4] We investigate the post-1990 period in more detail next, but this finding pointing to a negligible role in trade post-1990 appears rather robust across all different specifications of the gravity regression model estimated in this paper.

5.3 Quantifying the Extensive and Intensive Margins of Spoken Language (1995-2010)

When we look at the evidence in the post-1990 period, we see that indirect communication matters and that it is statistically significant for both dummy variables involving a common shared language that is non-native to at least one of the trading partners. These results are reported in Table 2. The estimated partial effect of communication through a language that is nonnative to both trading partners appears small, but this nonetheless showcases a significant reversal from the early period between 1970 and 1990. We think that being able to communicate on a third language (often English) is progressively becoming a more important venue to facilitate trade. In the early part of our sample, that effect was not present partly because foreign

[4] When we take out the variable that accounts for the free-trade agreement in common between the countries, we still find that the common currency variable is insignificant in the post-1990 period.

language education may have been determined partly by non-economic considerations and perhaps the number of individual who knew a second language was not so significant back then, but over time it has become more important. Noticeably, however, the major impact over the period between 1995 and 2010 comes from sharing a language that is nonnative to one country but native to another. The estimated partial effect is not only statistically significant, but also quite sizeable. As a way of putting into context the impact that we have estimated, notice that the estimated coefficient is almost half that of the common native language (a bit over 45 percent to be more precise) and is as large as the effect of sharing a common official language. In other words, it has become as important as the official status of a language in facilitating trade. Our interpretation of these findings is that over time the ability to speak a second language has been particularly useful for countries whose native language is not widely spoken to embed themselves into trading networks where one of the major trading languages can be useful. For instance, it is not surprising that the most widely spoken language in Brazil that is nonnative to the country would be Spanish and that Brazil trades more intensely with Spanish-speaking countries as a result of its geographical configuration or that Japan, whose nonnative language is English, tends to trade more intensely with English-speaking countries.

5.4 The Role of Diversity and Literacy in Spoken Language (1995-2010)

We have more information about the share of spoken languages to characterize the intensity with which native and nonnative languages are utilized for the post-1990 period based on the available data from Ethnologue. We also have additional information on country characteristics relating to the literacy rates of the education languages (the official and native languages) as well as the internal linguistic diversity of each country. We also have detailed information regarding the similarity across native languages that are not shared in common across countries. So far, we have been able to assess whether speaking a common second language or not would have any effect on trade. With this additional data sources and the greater detail they bring about the linguistic characteristics of each bilateral pair, we can assess more remarkable questions for the post-1990 period. We emphasize those interesting trade-offs here investigating in particular four specific questions: controlling for official status (a) Does it matter the distinction between L1 and L2 use in each country? Does it matter the fraction of the population that can speak a second language? (b) Does it matter the literacy of the population on native languages for written communication? (c) Does it matter the degree of linguistic diversity of native languages of the country to promote international trade instead? (d) Does it matter the degree of linguistic similarity between the not-in-common native languages in each bilateral trade? Is there a trade-off between linguistic similarity facilitating trade and making the impact of a second language less relevant?

Several hypotheses could be discussed with respect to these questions. On the one hand, assessing the intensity with which a common (second or otherwise) language is shared across countries is important to determine how relevant is the degree and the intensity with which native and nonnative languages are spoken, in order to account for the trade patterns that arise in the data. If we control for all other factors distinguishing two countries, which impact on trade with the US does it have that a given country (e.g., the Netherlands) has a high proportion of second language learners of English, as compared with another one with a smaller proportion of English learners (e.g., Japan)?

In Table 3, we observe that, when we account for the intensity of use of the second language spoken by the population of two countries (where the language is nonnative), we find a weak statistically significant negative effect. We find the opposite effect with the corresponding dummy variable. We interpret this as evidence indicating that the overall impact on trade from this channel is probably negligible and does not contribute to facilitate trade at this point. Notice that all other covariates included in the model remain almost unaltered across model specifications, and that, as expected for the post-1990 period, we do not observe a significant effect from the currency union variable, while trade agreements are both large and statistically significant in our sample.

On the other hand, another question remains open, which is whether country characteristics on linguistics data affect the role and significance of the partial effects estimated for the role of communication through second languages. It is important to note that the country characteristics of literacy and diversity have been used in the past in Melitz (2008), but were not considered before in the context of exploring the role of second languages. We also find that these country characteristics have an effect on their own that cannot be picked up by the country fixed effects in the model specification used in this paper, which deserves further exploration. These country characteristics were not included in the work of Melitz and Toubal (2014), but they are quite important in that they contribute to alter significantly the estimated effect of the second language variables constructed, as well as the effectiveness of an overall common spoken language in order to account for the observed ex post trade between bilateral country pairs in our sample for the post-1990 period.

Our estimates indicate that there is a significant attenuation of the effect of sharing a common native language (based on L1 and L2 speakers), that the direct effect of communication through a nonnative language is rather negligible and statistically not different from zero, and that the effect of sharing a common nonnative language between the two countries tends to result in a smaller trade trend among them. There is strong evidence as well, of a significant positive interaction between diversity and the variables of nonnative and native-nonnative language. This interaction would suggest that there is an important connection between the two variables, such that the countries with more diversity are also more likely to trade with other countries, because they may be able to speak the language of the other country. It could be that a native language in another country is spoken in this specific country either as a second language or a native (but unofficial) language. This diversity increases, thus, the number of potential trading partners of a country.

The effect of common official language almost disappears as can be seen in Tables 3 and 4, as soon as we include variables that capture the potential interactions between the spoken language disaggregated variables and literacy in particular. This prompts us to re-evaluate our understanding of the role that official languages play in facilitating trade. We would argue that official languages tend to be the languages of education. Therefore, literacy and the interaction between literacy and spoken languages can offer a better characterization of the channels through which language education affects trade than the common official dummy by itself did. In this regard, we find this to be a very novel result.

5.5 The Role of Language Similarity (1995-2010)

Using an alternative representation of native language, Table 4 allows for a more gradient representation of the similarity among languages (LP1 or LP2). Melitz and Toubal (2014) found these two variables to be highly reliable measures of international trade. The differences reported between the two native language similarity measures are small, but still positive and statistically significant. Our findings reveal that the relatively greater trade facilitation of indirect communication between non-native and native speakers rather than two non-native speakers is also robust in this case. We observe that linguistic similarity has a sizeable and significant effect, but the effect of non-native spoken languages is attenuated by at least 1/3, indicating that there are different competing channels that explain these trade effects—the role of second-languages and language similarity. This suggests that the multiple channels through which language affects trade are not fully incorporated in the model. Future studies should focus on better understanding the roles that both similarity and second-languages have on international trade.

Table 1: Common language Regressand: log of bilateral trade over product of GDP

VARIABLES	LABELS	(1) y 1970-2010	(2) y 1970-1990	(3) y 1995-2010	(4) y 1970-2010	(5) y 1970-1990	(6) y 1995-2010	(7) y 1970-2010	(8) y 1970-1990	(9) y 1995-2010	(10) y 1970-2010	(11) y 1970-1990	(12) y 1995-2010
y	year	.	.	.	.	.	.	.	.	.	.	.	.
		(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)
col	common official language	0.420***	0.340***	0.509***	0.284***	0.298***	0.318***	0.248***	0.261***	0.285***	0.241***	0.302***	0.245***
		(10.885)	(6.801)	(11.859)	(6.293)	(5.024)	(6.398)	(6.030)	(4.875)	(6.238)	(5.209)	(4.934)	(4.829)
cnldummy	dummy equals 1 if cnl>0, 0 otherwise				0.333***	0.098	0.486***						
					(6.016)	(1.316)	(8.234)						
csldummy	dummy equals 1 if csl>0, 0 otherwise							0.347***	0.159***	0.452***			
								(11.038)	(3.640)	(13.104)			
cslnativedummy	dummy equals 1 if cslnative>0, 0 otherwise										0.416***	0.099	0.620***
											(7.053)	(1.232)	(9.915)
cslnonnativedummy	dummy equals 1 if cslnonnative>0, 0 otherwise										-0.056	-0.204***	0.069*
											(-1.549)	(-4.113)	(1.709)
cslnativenonnativedummy	dummy equals 1 if cslnativenonnative>0, 0 otherwise										0.142***	-0.054	0.278***
											(4.324)	(-1.218)	(7.742)
logdist	geodesic distance in logs	-1.402***	-1.316***	-1.487***	-1.376***	-1.308***	-1.450***	-1.392***	-1.310***	-1.478***	-1.385***	-1.322***	-1.460***
		(-74.258)	(-52.978)	(-71.561)	(-71.361)	(-51.661)	(-68.263)	(-73.597)	(-52.600)	(-71.087)	(-70.987)	(-51.566)	(-68.031)
comborder	common border (0,1) dummy	0.449***	0.324***	0.583***	0.414***	0.316***	0.520***	0.417***	0.314***	0.533***	0.385***	0.323***	0.471***
		(5.151)	(3.101)	(5.891)	(4.717)	(3.021)	(5.222)	(4.847)	(3.025)	(5.448)	(4.369)	(3.075)	(4.694)
comcur	country pair in currency union or currency board	0.492***	1.150***	-0.010	0.509***	1.149***	0.033	0.482***	1.131***	0.005	0.528***	1.179***	0.028
		(4.521)	(8.143)	(-0.081)	(4.647)	(8.117)	(0.267)	(4.453)	(8.022)	(0.041)	(4.791)	(8.320)	(0.220)
rta	country pair in free trade agreement	0.418***	-0.080	0.426***	0.411***	-0.084	0.419***	0.419***	-0.047	0.406***	0.410***	-0.090	0.403***
		(9.808)	(-0.772)	(9.914)	(9.720)	(-0.818)	(9.860)	(9.925)	(-0.456)	(9.596)	(9.661)	(-0.874)	(9.473)
comcol	country pair with a common colonizer post 1945	0.810***	0.596***	0.950***	0.833***	0.609***	0.964***	0.774***	0.583***	0.894***	0.842***	0.650***	0.952***
		(16.059)	(8.617)	(17.018)	(16.410)	(8.724)	(17.326)	(15.336)	(8.413)	(16.132)	(16.498)	(9.209)	(17.089)
col45	country pair in colonial relationship post 1945	1.185***	1.065***	1.271***	1.241***	1.089***	1.323***	1.208***	1.081***	1.299***	1.210***	1.050***	1.314***
		(8.900)	(6.952)	(8.876)	(9.288)	(7.027)	(9.286)	(9.148)	(7.078)	(9.149)	(9.026)	(6.793)	(9.178)
colony	country pair ever in a colonial relationship	0.602***	0.818***	0.361***	0.556***	0.803***	0.299***	0.667***	0.852***	0.433***	0.562***	0.776***	0.316***
		(6.363)	(7.149)	(3.485)	(5.809)	(6.959)	(2.873)	(7.038)	(7.441)	(4.179)	(5.801)	(6.731)	(2.984)
curcol	country pair in current colonial relationship	-0.232	-0.069	-1.693*	-0.313	-0.090	-1.914*	-0.205	-0.061	-1.645	-0.300	-0.073	-1.889*
		(-0.535)	(-0.219)	(-1.722)	(-0.702)	(-0.287)	(-1.878)	(-0.460)	(-0.191)	(-1.592)	(-0.668)	(-0.235)	(-1.797)
comleg	country pair share common legal origin	0.225***	0.205***	0.230***	0.221***	0.204***	0.223***	0.188***	0.192***	0.173***	0.208***	0.211***	0.188***
		(8.507)	(5.413)	(8.053)	(8.348)	(5.396)	(7.808)	(7.055)	(5.048)	(6.054)	(7.743)	(5.560)	(6.508)
relprox	country pair share same religion	0.480***	0.337***	0.587***	0.411***	0.316***	0.490***	0.377***	0.285***	0.461***	0.369***	0.323***	0.420***
		(9.148)	(4.552)	(10.225)	(7.731)	(4.212)	(8.433)	(7.092)	(3.807)	(7.916)	(6.807)	(4.218)	(7.099)
war	years at war	-1.496***	-1.591***	-1.144***	-1.471***	-1.583***	-1.117***	-1.474***	-1.580***	-1.126***	-1.489***	-1.572***	-1.155***
		(-11.818)	(-10.376)	(-7.472)	(-11.645)	(-10.316)	(-7.383)	(-11.735)	(-10.315)	(-7.466)	(-11.805)	(-10.254)	(-7.618)
Observations		131,322	55,895	75,427	131,322	55,895	75,427	131,322	55,895	75,427	131,322	55,895	75,427
R-squared		0.724	0.696	0.740	0.724	0.696	0.740	0.725	0.697	0.741	0.724	0.697	0.741
mss		1.422e+06	507682	906580	1.422e+06	507703	907311	1.423e+06	507815	908130	1.423e+06	507861	907800
rss		542790	221328	319158	542204	221307	318427	541235	221195	317608	541860	221148	317938
r2		0.724	0.696	0.740	0.724	0.696	0.740	0.725	0.697	0.741	0.724	0.697	0.741
N_clust		24753	15106	23029	24753	15106	23029	24753	15106	23029	24753	15106	23029
df_r		2803	1465	1350	2804	1466	1351	2804	1466	1351	2806	1468	1353

Robust t-statistics in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 2: Common language by population share Regressand: log of real bilateral trade (1995-2010)

VARIABLES	LABELS	(1) y dummy	(2) y dummy	(3) y dummy	(4) y index	(5) y index	(6) y index	(7) y index
y	year	.	.	.	.	.	.	.
		(.)	(.)	(.)	(.)	(.)	(.)	(.)
col	common official language	0.318***	0.285***	0.245***	0.324***	0.282***	0.303***	0.243***
		(6.398)	(6.238)	(4.829)	(6.429)	(5.448)	(5.962)	(4.604)
cnl	common native language - based on L1 speakers only	0.486***			0.768***			
		(8.234)			(7.615)			
csl	common spoken language index - based on all speakers (L1+L2)		0.452***				0.558***	
			(13.104)				(7.729)	
cslnative	common spoken language index - all speakers - native-to-native			0.620***		0.677***		0.772***
				(9.915)		(8.150)		(8.992)
cslnonnative	common spoken language index-all spoken-nonnative-to-nonnative			0.069*				-0.304*
				(1.709)				(-1.699)
cslnativenonnative	common spoken language index - all spoken - native-to-nonnative			0.278***				0.397***
				(7.742)				(4.026)
logdist	geodesic distance in logs	-1.450***	-1.478***	-1.460***	-1.459***	-1.454***	-1.466***	-1.463***
		(-68.263)	(-71.087)	(-68.031)	(-68.444)	(-67.869)	(-69.397)	(-67.321)
comborder	common border (0,1) dummy	0.520***	0.533***	0.471***	0.573***	0.558***	0.547***	0.537***
		(5.222)	(5.448)	(4.694)	(5.788)	(5.626)	(5.507)	(5.408)
comcur	country pair in currency union or currency board	0.033	0.005	0.028	0.037	0.039	0.002	0.058
		(0.267)	(0.041)	(0.220)	(0.302)	(0.319)	(0.015)	(0.469)
rta	country pair in free trade agreement	0.419***	0.406***	0.403***	0.429***	0.425***	0.397***	0.432***
		(9.860)	(9.596)	(9.473)	(10.022)	(9.977)	(9.325)	(10.000)
comcol	country pair with a common colonizer post 1945	0.964***	0.894***	0.952***	0.998***	0.993***	0.982***	1.002***
		(17.326)	(16.132)	(17.089)	(17.723)	(17.743)	(17.599)	(17.919)
col45	country pair in colonial relationship post 1945	1.323***	1.299***	1.314***	1.354***	1.342***	1.320***	1.326***
		(9.286)	(9.149)	(9.178)	(9.422)	(9.369)	(9.258)	(9.261)
colony	country pair ever in a colonial relationship	0.299***	0.433***	0.316***	0.300***	0.287***	0.300***	0.276***
		(2.873)	(4.179)	(2.984)	(2.893)	(2.779)	(2.884)	(2.658)
curcol	country pair in current colonial relationship	-1.914*	-1.645	-1.889*	-1.953**	-1.864*	-1.777*	-1.792*
		(-1.878)	(-1.592)	(-1.797)	(-2.055)	(-1.936)	(-1.838)	(-1.838)
comleg	country pair share common legal origin	0.223***	0.173***	0.188***	0.227***	0.220***	0.214***	0.216***
		(7.808)	(6.054)	(6.508)	(7.954)	(7.693)	(7.508)	(7.547)
relprox	country pair share same religion	0.490***	0.461***	0.420***	0.498***	0.502***	0.489***	0.467***
		(8.433)	(7.916)	(7.099)	(8.522)	(8.646)	(8.312)	(7.961)
war	years at war	-1.117***	-1.126***	-1.155***	-1.115***	-1.111***	-1.153***	-1.123***
		(-7.383)	(-7.466)	(-7.618)	(-7.301)	(-7.318)	(-7.452)	(-7.339)
Observations		75,427	75,427	75,427	75,427	75,427	75,427	75,427
R-squared		0.740	0.741	0.741	0.740	0.740	0.740	0.740
mss		907311	908130	907800	907098	907200	907117	907406
rss		318427	317608	317938	318640	318538	318621	318332
r2		0.740	0.741	0.741	0.740	0.740	0.740	0.740
N_clust		23029	23029	23029	23029	23029	23029	23029
df_r		1351	1351	1353	1351	1351	1351	1353

Robust t-statistics in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3: Common language effect with diversity and literacy Regressand: log of real bilateral trade (1995-2010)

VARIABLES	LABELS	(1) y index	(2) y index	(3) y index	(4) y index	(5) y index	(6) y index	(7) y index	(8) y index	(9) y index	(10) y index
y	year	.	.	.	.	.	.	.	.	.	.
		(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)
col	common official language	0.303***	0.308***	0.313***	0.318***	0.243***	0.241***	0.241***	0.242***	0.074	0.073
		(5.962)	(6.001)	(6.193)	(6.220)	(4.604)	(4.535)	(4.571)	(4.571)	(1.330)	(1.317)
csl	common spoken language index - based on all speakers (L1+L2)	0.558***	0.541***	0.562***	0.543***						
		(7.729)	(7.389)	(7.787)	(7.421)						
cslnative	common spoken language index - all speakers - native-to-native					0.772***	0.784***	0.825***	0.823***	2.039***	2.035***
						(8.992)	(9.079)	(9.582)	(9.506)	(7.016)	(7.399)
cslnonnative	common spoken language index-all spoken-nonnative-to-nonnative					-0.304*	-0.507***	-0.392**	-0.589***	5.896***	5.718***
						(-1.699)	(-2.721)	(-2.190)	(-3.154)	(6.388)	(7.713)
cslnativenonnative	common spoken language index - all spoken - native-to-nonnative					0.397***	0.339***	0.335***	0.292***	1.542***	1.536***
						(4.026)	(3.376)	(3.397)	(2.908)	(3.216)	(3.209)
cslnative_int	cslnative_int									-0.021	
										(-0.052)	
cslnonnative_int	cslnonnative_int									-0.362	
										(-0.338)	
cslnativenonnative_int	cslnativenonnative_int									1.838***	1.862***
										(3.150)	(3.220)
cslnative_int2	cslnative_int2									-1.329***	-1.325***
										(-4.102)	(-4.175)
cslnonnative_int2	cslnonnative_int2									-7.662***	-7.501***
										(-7.985)	(-9.258)
cslnativenonnative_int2	cslnativenonnative_int2									-1.670***	-1.666***
										(-3.316)	(-3.311)
diversity	Product of domestic linguistic diversity index of each country		1.073***		0.852***		1.081***		0.844***	0.889***	0.881***
			(8.310)		(6.459)		(8.364)		(6.400)	(6.620)	(6.651)
literacy	Product of the literacy rate of each country in the pair			2.537***	2.096***			2.704***	2.276***	2.569***	2.569***
				(8.795)	(7.070)			(9.316)	(7.642)	(8.477)	(8.476)
logdist	geodesic distance in logs	-1.466***	-1.441***	-1.458***	-1.438***	-1.463***	-1.438***	-1.451***	-1.433***	-1.448***	-1.448***
		(-69.397)	(-67.082)	(-68.859)	(-66.946)	(-67.321)	(-64.872)	(-66.599)	(-64.628)	(-65.326)	(-65.312)
comborder	common border (0,1) dummy	0.547***	0.561***	0.509***	0.529***	0.537***	0.552***	0.496***	0.517***	0.499***	0.498***
		(5.507)	(5.664)	(5.139)	(5.348)	(5.408)	(5.583)	(5.012)	(5.237)	(5.101)	(5.111)
comcur	country pair in currency union or currency board	0.002	-0.050	-0.164	-0.173	0.058	0.019	-0.110	-0.108	-0.137	-0.144
		(0.015)	(-0.403)	(-1.324)	(-1.401)	(0.469)	(0.154)	(-0.894)	(-0.880)	(-1.117)	(-1.195)
rta	country pair in free trade agreement	0.397***	0.373***	0.341***	0.332***	0.432***	0.410***	0.376***	0.368***	0.443***	0.444***
		(9.325)	(8.561)	(7.959)	(7.597)	(10.000)	(9.313)	(8.686)	(8.369)	(10.062)	(10.080)
comcol	country pair with a common colonizer post 1945	0.982***	0.988***	0.936***	0.951***	1.002***	1.013***	0.956***	0.975***	0.902***	0.902***
		(17.599)	(17.609)	(16.666)	(16.843)	(17.919)	(18.013)	(17.023)	(17.266)	(15.643)	(15.679)
col45	country pair in colonial relationship post 1945	1.320***	1.352***	1.340***	1.359***	1.326***	1.369***	1.353***	1.381***	1.369***	1.369***
		(9.258)	(9.408)	(9.386)	(9.448)	(9.261)	(9.474)	(9.412)	(9.535)	(9.431)	(9.442)
colony	country pair ever in a colonial relationship	0.300***	0.288***	0.306***	0.295***	0.276***	0.253**	0.277***	0.255**	0.305***	0.305***
		(2.884)	(2.734)	(2.942)	(2.802)	(2.658)	(2.404)	(2.667)	(2.431)	(2.916)	(2.920)
curcol	country pair in current colonial relationship	-1.777*	-1.820*	-1.812*	-1.836*	-1.792*	-1.835*	-1.847*	-1.863*	-1.802*	-1.800*
		(-1.838)	(-1.905)	(-1.882)	(-1.923)	(-1.838)	(-1.899)	(-1.903)	(-1.930)	(-1.758)	(-1.756)
comleg	country pair share common legal origin	0.214***	0.211***	0.215***	0.210***	0.216***	0.212***	0.216***	0.211***	0.217***	0.217***
		(7.508)	(7.375)	(7.537)	(7.355)	(7.547)	(7.388)	(7.583)	(7.368)	(7.574)	(7.600)
relprox	country pair share same religion	0.489***	0.456***	0.404***	0.393***	0.467***	0.430***	0.374***	0.358***	0.341***	0.342***
		(8.312)	(7.664)	(6.804)	(6.544)	(7.961)	(7.232)	(6.302)	(5.970)	(5.626)	(5.652)
war	years at war	-1.153***	-1.185***	-1.166***	-1.193***	-1.123***	-1.145***	-1.129***	-1.148***	-1.095***	-1.094***
		(-7.452)	(-7.728)	(-7.663)	(-7.869)	(-7.339)	(-7.571)	(-7.536)	(-7.697)	(-7.444)	(-7.436)
Observations		75,427	73,921	75,427	73,921	75,427	73,921	75,427	73,921	73,921	73,921
R-squared		0.740	0.741	0.741	0.741	0.740	0.741	0.741	0.742	0.743	0.743
mss		907117	889109	907920	889625	907406	889497	908308	890099	891272	891270
rss		318621	310756	317818	310240	318332	310369	317430	309766	308594	308595
r2		0.740	0.741	0.741	0.741	0.740	0.741	0.741	0.742	0.743	0.743
N_clust		23029	22531	23029	22531	23029	22531	23029	22531	22531	22531
df_r		1351	1336	1352	1337	1353	1338	1354	1339	1345	1343

Robust t-statistics in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 4: Linguistic proximity Regressand: log of real bilateral trade (1995-2010)

VARIABLES	LABELS	(1) y no-lp	(2) y no-lp	(3) y no-lp	(4) y lp1	(5) y lp1	(6) y lp1	(7) y lp1	(8) y lp2	(9) y lp2	(10) y lp2	(11) y lp2
y	year	.	.	.	.	.	.	.	.	.	.	.
		(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)	(.)
col	common official language	0.242***	0.074	0.073	0.257***	0.110*	0.099*	0.100*	0.262***	0.112*	0.105*	0.112*
		(4.571)	(1.330)	(1.317)	(4.664)	(1.907)	(1.687)	(1.735)	(4.752)	(1.946)	(1.798)	(1.961)
lp	language proximity				0.525***	0.536***	0.973***	0.946***	0.955***	1.037***	1.908***	1.853***
					(4.733)	(4.688)	(7.315)	(7.452)	(5.280)	(5.668)	(8.169)	(8.359)
cslnative	common spoken language index - all speakers - native-to-native	0.823***	2.039***	2.035***	0.906***	1.838***	1.788***	1.692***	0.932***	1.962***	1.998***	1.892***
		(9.506)	(7.016)	(7.399)	(10.191)	(6.208)	(6.028)	(6.115)	(10.376)	(6.609)	(6.714)	(6.849)
cslnonnative	common spoken language index-all spoken-nonnative-to-nonnative	-0.589***	5.896***	5.718***	-0.473**	5.564***	4.632***	4.476***	-0.493**	5.651***	5.007***	4.664***
		(-3.154)	(6.388)	(7.713)	(-2.433)	(5.893)	(4.860)	(5.859)	(-2.538)	(5.988)	(5.273)	(6.109)
cslnativenonnative	common spoken language index - all spoken - native-to-nonnative	0.292***	1.542***	1.536***	0.250**	1.129**	0.738	0.718***	0.223**	1.171**	0.953*	0.642***
		(2.908)	(3.216)	(3.209)	(2.337)	(2.228)	(1.452)	(4.346)	(2.079)	(2.308)	(1.880)	(3.633)
cslnative_int	cslnative*diversity		-0.021			-0.182	-0.360			-0.132	-0.276	
			(-0.052)			(-0.432)	(-0.846)			(-0.313)	(-0.648)	
cslnonnative_int	cslnonnative*diversity		-0.362			-0.335	-0.352			-0.438	-0.650	
			(-0.338)			(-0.289)	(-0.309)			(-0.378)	(-0.567)	
cslnativenonnative_int	cslnativenonnative*diversity		1.838***	1.862***		2.196***	1.737**	1.820***		2.199***	1.739**	1.956***
			(3.150)	(3.220)		(3.077)	(2.435)	(2.722)		(3.080)	(2.434)	(2.910)
cslnative_int2	cslnative*literacy		-1.329***	-1.325***		-0.970***	-0.791**	-0.734**		-1.086***	-0.970***	-0.911***
			(-4.102)	(-4.175)		(-2.930)	(-2.366)	(-2.267)		(-3.300)	(-2.906)	(-2.839)
cslnonnative_int2	cslnonnative*literacy		-7.662***	-7.501***		-7.210***	-5.102***	-4.967***		-7.330***	-5.623***	-5.298***
			(-7.985)	(-9.258)		(-7.305)	(-4.946)	(-5.664)		(-7.431)	(-5.458)	(-5.995)
cslnativenonnative_int2	cslnativenonnative*literacy		-1.670***	-1.666***		-1.265**	-0.010			-1.356**	-0.356	
			(-3.316)	(-3.311)		(-2.369)	(-0.018)			(-2.534)	(-0.619)	
cslnative_lpint	cslnative*lp						-0.166				-0.536	
							(-0.271)				(-0.629)	
cslnonnative_lpint	cslnonnative*lp						-5.363***	-5.326***			-6.714***	-6.675***
							(-6.791)	(-6.757)			(-5.608)	(-5.569)
cslnativenonnative_lpint	cslnativenonnative*lp						-2.675***	-2.640***			-3.738***	-3.800***
							(-5.415)	(-5.814)			(-4.550)	(-5.064)
diversity	Product of domestic linguistic diversity index of each country	0.844***	0.889***	0.881***	0.646***	0.708***	0.683***	0.675***	0.659***	0.722***	0.698***	0.681***
		(6.400)	(6.620)	(6.651)	(4.751)	(5.107)	(4.932)	(4.943)	(4.869)	(5.242)	(5.063)	(5.016)
literacy	Product of the literacy rate of each country in the pair	2.276***	2.569***	2.569***	2.261***	2.495***	2.358***	2.364***	2.224***	2.460***	2.319***	2.306***
		(7.642)	(8.477)	(8.476)	(7.508)	(8.165)	(7.727)	(7.819)	(7.371)	(8.040)	(7.572)	(7.602)
logdist	geodesic distance in logs	-1.433***	-1.448***	-1.448***	-1.426***	-1.437***	-1.441***	-1.442***	-1.424***	-1.434***	-1.439***	-1.439***
		(-64.628)	(-65.326)	(-65.312)	(-63.145)	(-63.617)	(-63.764)	(-63.813)	(-62.663)	(-63.134)	(-63.235)	(-63.299)
comborder	common border (0,1) dummy	0.517***	0.499***	0.498***	0.543***	0.530***	0.522***	0.514***	0.545***	0.531***	0.529***	0.519***
		(5.237)	(5.101)	(5.111)	(5.511)	(5.405)	(5.336)	(5.277)	(5.539)	(5.422)	(5.395)	(5.335)
comcur	country pair in currency union or currency board	-0.108	-0.137	-0.144	0.057	-0.005	0.034	0.027	0.051	-0.010	0.005	-0.008
		(-0.880)	(-1.117)	(-1.195)	(0.432)	(-0.041)	(0.257)	(0.208)	(0.387)	(-0.077)	(0.035)	(-0.060)
rta	country pair in free trade agreement	0.368***	0.443***	0.444***	0.384***	0.454***	0.455***	0.455***	0.383***	0.454***	0.447***	0.446***
		(8.369)	(10.062)	(10.080)	(8.580)	(10.123)	(10.195)	(10.192)	(8.545)	(10.130)	(10.022)	(9.992)
comcol	country pair with a common colonizer post 1945	0.975***	0.902***	0.902***	0.954***	0.885***	0.885***	0.882***	0.959***	0.890***	0.895***	0.890***
		(17.266)	(15.643)	(15.679)	(16.741)	(15.207)	(15.154)	(15.195)	(16.886)	(15.341)	(15.338)	(15.409)
col45	country pair in colonial relationship post 1945	1.381***	1.369***	1.369***	1.345***	1.341***	1.311***	1.307***	1.347***	1.341***	1.324***	1.327***
		(9.535)	(9.431)	(9.442)	(9.148)	(9.068)	(8.723)	(8.738)	(9.130)	(9.036)	(8.810)	(8.859)
colony	country pair ever in a colonial relationship	0.255**	0.305***	0.305***	0.260**	0.294***	0.292***	0.295***	0.251**	0.291***	0.282***	0.281***
		(2.431)	(2.916)	(2.920)	(2.468)	(2.793)	(2.703)	(2.741)	(2.384)	(2.759)	(2.621)	(2.610)
curcol	country pair in current colonial relationship	-1.863*	-1.802*	-1.800*	-1.812*	-1.769*	-1.854*	-1.839*	-1.820*	-1.769*	-1.816*	-1.805*
		(-1.930)	(-1.758)	(-1.756)	(-1.873)	(-1.738)	(-1.868)	(-1.849)	(-1.883)	(-1.733)	(-1.833)	(-1.816)
comleg	country pair share common legal origin	0.211***	0.217***	0.217***	0.207***	0.212***	0.210***	0.211***	0.197***	0.201***	0.197***	0.198***
		(7.368)	(7.574)	(7.600)	(7.100)	(7.244)	(7.178)	(7.234)	(6.680)	(6.806)	(6.664)	(6.736)
relprox	country pair share same religion	0.358***	0.341***	0.342***	0.313***	0.299***	0.272***	0.276***	0.319***	0.299***	0.264***	0.270***
		(5.970)	(5.626)	(5.652)	(5.024)	(4.768)	(4.321)	(4.416)	(5.159)	(4.808)	(4.210)	(4.340)
war	years at war	-1.148***	-1.095***	-1.094***	-1.246***	-1.182***	-1.152***	-1.151***	-1.240***	-1.177***	-1.155***	-1.153***
		(-7.697)	(-7.444)	(-7.436)	(-8.084)	(-7.800)	(-7.903)	(-7.897)	(-8.052)	(-7.773)	(-7.849)	(-7.839)
Observations		73,921	73,921	73,921	71,865	71,865	71,865	71,865	71,865	71,865	71,865	71,865
R-squared		0.742	0.743	0.743	0.740	0.741	0.741	0.741	0.740	0.741	0.741	0.741
mss		890099	891272	891270	860457	861393	861957	861949	860494	861467	861873	861859
rss		309766	308594	308595	302575	301639	301075	301083	302538	301565	301159	301173
r2		0.742	0.743	0.743	0.740	0.741	0.741	0.741	0.740	0.741	0.741	0.741
N_clust		22531	22531	22531	21643	21643	21643	21643	21643	21643	21643	21643
df_r		1339	1345	1343	1326	1332	1335	1331	1326	1332	1335	1331

Robust t-statistics in parentheses

*** p<0.01, ** p<0.05, * p<0.1

6 Conclusion

The newly created variable of non-native spoken languages provides a more detailed measure of the linguistic factors that can affect trade. Our estimates indicate that the knowledge of other languages, specifically widely spoken languages such as English, significantly promotes bilateral trade. Other studies should expand on the evidence presented here to further disentangle the competing effects of non-native languages and language similarity on trade.

The increased effect on trade facilitation associated with the language variables suggests that indirect communication has become more important over the past two decades. It seems to be the case that the effect of language similarity was stronger as the beginning of our data sample, while in the last years, the role of second language acquisition has strengthened.

References
Papers and Publications

Anderson, James E. (1979): "A Theoretical Foundation for the Gravity Equation." *American Economic Review* 69(1):106-116.

Anderson, James E. and Eric van Wincoop (2003): "Gravity with Gravitas: A Solution to the Border Puzzle." *American Economic Review* 93(1):170-192.

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

Boisso, D. and M. Ferrantino (1997): "Economic Distance, Cultural Distance, and Openness in International Trade: Empirical Puzzles." *Journal of Economic Integration*.

CIA (2011): "Central Intelligence Agency: The World Factbook." (https://www.cia.gov/library/publications/the-world-factbook/).

Crystal, David. (2003): *English as a Global Language*. Second Edition ed. Cambridge, United Kingdom: Cambridge University Press.

Crystal, David. (2003): *The Cambridge Encyclopedia of the English Language*. Second Edition ed. Cambridge, United Kingdom: Cambridge University Press.

Deardorff, A. (1998): "Determinants of Bilateral Trade: Does Gravity Work in a Neoclassical World?" in *The Regionalization of the World Economy*, edited by J.A. Frankel. Chicago: University of Chicago Press.

Eaton, J. and S. Kortum (2002): "Technology, Geography, and Trade." *Econometrica* 70:1741-1799.

Egger, Peter H. and Andrea Lassman (2012): "The Language Effect in International Trade: A Meta-Analysis." *Economics Letters* 116:221-224.

Ginsburgh, Victor and Shlomo Weber. (2011): *How Many Languages Do We Need? The Economics of Linguistic Diversity*. Princeton, New Jersey: Princeton University Press.

Greenberg, J. H. (1956): "The Measurement of Linguistic Diversity." *LANGUAGE (Journal of the Linguistic Society of America)* 32:109-115.

Head, Keith and Thierry Mayer (2015): "Gravity Equations: Toolkit, Cookbook, Workhorse." Pp. 131-196 in *Handbook of International Economics*, edited by Gita Gopinath, Elhanan Helpman, and Kenneth Rogoff. Oxford: U.K.: North-Holland: Elsevier.

Helliwell, John F. (1999): "Language and Trade." Pp. 5-30 in *Exploring the Economics of Language*, edited by Albert Breton. Canadian Heritage.

Holman, Eric W. (2011): "Programs for Calculating ASJP Distance Matrices (version 2.1)." *Mimeo*. (http://email.eva.mpg.de/~wichmann/software.htm).

Lewis, M. P. (.). (2009): *Ethnologue: Languages of the World*. 16th ed. Dallas, TX: SIL International. Online. (http://www.ethnologue.com/).

Lieberson, S. (1981): *Language Diversity and Language Contact: Essays*. Stanford University Press.

Lohmann, Johannes (2011): "Do Language Barriers Affect Trade?" *Economics Letters* 110:159-162.

Melitz, Jacques (2008): "Language and Foreign Trade." *European Economic Review* 52:667-699.

Melitz, Jacques and Farid Toubal (2014): "Native Language, Spoken Language, Translation and Trade." *Journal of International Economics* 93:351-363.

Pöyhönen, P. (1963): "A Tentative Model for the Volume of Trade Between Countries." *Weltwertschaftliches Archiv*.
Rose, A.K. and E. van Wincoop (2001): "National Money as a Barrier to Trade: The Real Case for Currency Union." *American Economic Review* 91:386-390.
Santos Silva, JMC and S. Tenreyro (2006): "The Log of Gravity." *Review of Economics and Statistics* 88:641-658.
Tinbergen, Jan (1962): "An Analysis of World Trade Flows." in *Shaping the World Economy*, edited by Jan Tinbergen. New York, NY: Twentieth Century Fund.

Websites

Center for International Comparisons, University of Pennsylvania – Penn World Tables: https://pwt.sas.upenn.edu/
Centre d'Études Prospectives et d'Informations Internationales (CEPII): http://www.cepii.org/CEPII/en/bdd_modele/bdd.asp
U.N. COMTRADE Database: http://comtrade.un.org/
World Trade Organization (WTO)'s Regional Trade Agreement Gateway: http://rtais.wto.org/UI/PublicAllRTAList.aspx
CIA (Central Intelligence Agency) - The World Factbook: https://www.cia.gov/library/publications/the-world-factbook/index.html
SIL International - Ethnologue, Languages of the World: http://www.ethnologue.com/
The Max Planck Institute's Automated Similarity Judgment Program (ASJP): http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm

Appendix A. Data Sources and Definitions

List of Variables

Log real bilateral trade. The average annual bilateral trade flows computed with import data whenever available between countries *i* and *j* in year *t* in US $, deflated with the U.S. CPI (2005=100).

Common native language (CNL). The common native language dummy, a binary variable for whether both countries share a common primary or native tongue amongst the 42 native languages.

Common official language (COL). The common official language dummy, a binary variable for whether both countries share an official language amongst the 42 native languages.

Common spoken language index (CSL). The common spoken language index, increasing with the number of common languages (native or non-native) spoken by at least 4 percent of the population in a minimum of 2 countries and increasing with the likelihood that two individuals randomly selected in each trading partner are able to communicate in a given shared spoken (native or non-native) language. We construct also three sub-indexes depending on whether the spoken language is native to both countries in the pair, to only one, or to none.

Native language proximity (tree). Calculates the linguistic distance of two native languages looking at their position in the language family trees, and aggregates over all according to the share of native speakers. Assigns the value of 1 if at least 20 percent of the population of both countries speak a common native language.

Native language proximity (ASJP). Calculates the linguistic distance of two native languages looking at lexical aspects of the languages studied using the Automated Similarity Judgment Program (ASJP), and aggregates over all according to the share of native speakers. Assigns the value of 1 if at least 20 percent of the population of both countries speak a common native language.

Diversity. The product of the Greenberg's diversity index of both countries in the pair, as estimated by Lewis (2009). Greenberg's diversity index refers to the probability that any two people selected at random within a country would have different mother tongues (see Greenberg (1956), Lieberson (1981)).

Literacy. The product of the literacy rate of both countries in the pair.

Distance (log). The geodesic distance between country *i* and country *j*, calculated following the great circle (Haversine) formula between the cities that are the economic center of each country in the pair.

Product real GDP (log). Log product of the real GDP (in thousands), PPP-adjusted, of countries *i* and *j*.

Product real GDP per capita (log). Log-product of real GDP per capita, PPP-adjusted, of countries *i* and *j*.

Area (log). The product of the land areas of both countries, given in km^2.

Number landlocked (0,1,2). The number of landlocked countries (0, 1, 2) in each country-pair.

Common border. Binary variable that assigns the value 1 if the two countries share a border, 0 otherwise.

Common currency. Binary variable that assigns the value 1 if the two countries were part of a currency union or a currency board at the time, 0 otherwise.

Free-trade agreement. Binary variable that assigns the value 1 if the two countries were part of a free-trade agreement at the time, 0 otherwise.

Common colonizer post 1945. Binary variable that assigns the value 1 if the two countries share a common colonizer post 1945, 0 otherwise.*Ex-colonizer/colony (ever).* Binary variable that assigns the value of 1 if the two countries are or were ever in a colonial relationship, 0 otherwise.

Ex-colonizer/colony post 1945. Binary variable that assigns the value of 1 if the two countries were in a colonial relationship post 1945, 0 otherwise.

Ex-colonizer/colony (current). Binary variable that assigns the value of 1 if the two countries were in a colonial relationship at the time, 0 otherwise.

Common legal system. Binary variable that assigns the value of 1 if the two countries share Civil law, Common law or Muslim law, 0 otherwise.

Religious proximity. A common religion index constructed by adding up the products of population shares in both trading partners with the same religion (Buddhist, Catholic, Orthodox, Protestant, Hindu, Jewish, Shia, Sunni, and all others).

Years at war. Number of years at war for each country pair since 1823.

List of Common Languages

From Table 1 in Melitz and Toubal (2014), the 42 common native languages spoken as a first-language by at least 4 percent of the population in a minimum of 2 countries in our sample are:

Common official: Arabic, Bulgarian (plus Macedonian), Chinese, Danish (plus Icelandic), Dutch (plus Afrikaner), English, French, German, Greek, Italian, Malay, Farsi (plus Tajik), Portuguese, Romanian, Russian (plus Belarusian), Spanish, Swahili, Swedish, and Turkish (plus Turkmen and Azerbaijani).

Common non-official: Albanian, Armenian, Bengali, Bosnian, Croatian, Czech, Fang, Finnish, Fulfulde, Hausa, Hindi (plus Hindustani), Hungarian, Javanese, Lingala, Nepali, Pashto, Polish, Quechua, Serbian, Tamil, Ukrainian, Urdu, and Uzbek.

Data Sources

The bilateral trade data comes from the UN COMTRADE database, Bilateral trade observations (Revision 1), collected for years 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005, and 2010 (http://comtrade.un.org/). The U.S. CPI used to deflate bilateral trade is from Bureau of Labor Statistics (http://www.bls.gov/). Real GDP and real GDP per capita data were obtained from the Penn World Table (PWT) 8.0 (https://pwt.sas.upenn.edu/) completed with data from PWT 7.1, 6.3 and 5.6, as well as with data from the Conference Board's Total Economy Database (https://www.conference-board.org/data/economydatabase/).

The linguistic variables used and modified are constructed from Melitz and Toubal (2014), Lewis (2009) and CIA (2011). The geographic and distance measures (area, landlocked, geodesic distance, common border) are our own calculations using national sources. Data on common colonies (past and present), common currency, free-trade agreements, common legal system, common religion and year at war are obtained from Centre d'Études Prospectives et d'Informations Internationales (CEPII, http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=6), the World Trade Organization (http://www.wto.org/index.htm), and extended with various national sources.

Summary of the Data Sources

VARIABLE NAME	SHORT DESCRIPTION	TYPE (UNITS)
Dependent Variable		
TRADE	Bilateral trade flows	US dollars (in thousands) deflated with US CPI 2005=100, in logs
Main Regressors		
GDP	Product of the real GDP of each country in the pair	Log-product of each country's real GDP reported in PPP-adjusted values (in thousands)
GDPPC	Product of the real GDP per capita (to account for population density) of each country in the pair	Log-product of each country's real GDP per capita reported in PPP-adjusted values per population
DIST	Geodesic distance (Haversine formula) between main economic centers of each country	Great-circle distances calculated in km, in logs
Linguistic Distance		
OFFICIAL*	Country pair share common official language	Dummy Variable (0,1)
NATIVE	Common native language index (product of the fraction of native speakers summed over the two most-spoken native languages)	Index of commonality (≥ 0)
SPOKEN	Common spoken language index (product of the fraction of all speakers summed over all languages spoken by at least 4 percent of the population)	Index of commonality (≥ 0)
SPOKEN (EXC. NATIVE)*	Common language index for spoken language but not as a mother tongue (product of the fraction of all speakers summed over all languages spoken by at least 4 percent of the population)	Index of commonality (≥ 0)
LANGUAGE PROXIMITY (tree)	Linguistic Proximity 1. Similarity between common native languages (Tree method)	Distance measured by number of branches, interval 0-1
LANGUAGE PROXIMITY (ASJP)	Linguistic Proximity 2. Similarity between common native languages (ASJP method)	Distance measured by lexical similarity, interval 0-1
DIVERSITY	Product of the Greenberg's diversity index of each country in the pair	Greenberg's diversity index - Probability (0,1) that any two people randomly selected would have different mother tongues (Lieberson, 1981)
LITERACY	Product of the literacy rate of each country in the pair	Literacy rate - Share of the population (0,1) that is literate in its mother language
Geographic Distance (Other than Geodesic Distance)		
AREA	Product of the land area of each country in the pair	Log-product of each country's area reported in Km^2
LANDLOCKED	Number of countries without access to the sea in each pair	Dummy Variable (0,1,2)
COMBORDER	Country pair share a common border	Dummy Variable (0,1)
Geopolitical Distance (Institutions, History)		
COMCOL	Country pair with a common colonizer post 1945	Dummy Variable (0,1)
COLONY	Country pair ever in a colonial relationship	Dummy Variable (0,1)
COL45	Country pair in colonial relationship post 1945	Dummy Variable (0,1)
CURCOL	Country pair in current colonial relationship	Dummy Variable (0,1)
Geopolitical Distance (other)		
COMLEG	Country pair share common legal origin	Dummy Variable (0,1)
WAR	Number of years at war	Number of years
RELPROX	Religious proximity	Index interval 0-1
Economic Distance		
RTA	Free trade agreement in force	Dummy Variable (0,1)
COMCUR	Common currency	Dummy Variable (0,1)

* These variables are constructed for a subset of languages only: special status languages and english language. S*pecial status* includes Arabic, English, French, Portuguese and Spanish which are the languages spoken in the most countries.

Appendix B. Characteristics of the Data

Table B.1: Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max	Obs	Mean	Std. Dev.	Min	Max
year	1970-1990	1970-1990	1970-1990	1970-1990	1970-1990	1995-2010	1995-2010	1995-2010	1995-2010	1995-2010
logrealtrade	63980	14.85788	3.643635	1.614873	25.66701	90503	14.21686	4.11194	-0.1104947	26.56068
col	57327	0.1808746	0.3849176	0	1	78148	0.1471055	0.3542133	0	1
cnldummy	57327	0.1011914	0.3015846	0	1	78148	0.0894969	0.2854614	0	1
csldummy	57327	0.5500724	0.4974908	0	1	78148	0.463326	0.4986564	0	1
cslnativedummy	57327	0.1011914	0.3015846	0	1	78148	0.0894969	0.2854614	0	1
cslnonnativedummy	57327	0.2819439	0.4499501	0	1	78148	0.2245995	0.417321	0	1
cslnativenonnativedummy	57327	0.2217629	0.415436	0	1	78148	0.1926089	0.394351	0	1
lp1	54424	0.0799105	0.1536945	0	0.75	80567	0.0808121	0.146444	0	0.75
lp2	54424	0.0679772	0.085827	0	0.7417947	80567	0.070245	0.0846535	0	0.8326361
cnl	57327	0.0486935	0.1809175	0	0.99	78148	0.035929	0.1541636	0	0.99
csl	57327	0.1616011	0.2689443	0	1.6496	78148	0.1345544	0.2458959	0	1.6496
cslnative	57327	0.0687699	0.2267642	0	1	78148	0.05572	0.2023109	0	1
cslnonnative	57327	0.0343614	0.0918435	0	1.1915	78148	0.0281412	0.0834753	0	1.1915
cslnativenonnative	57327	0.0584698	0.1534384	0	0.89	78148	0.0506933	0.1405477	0	0.9456
diversity	62721	0.1716716	0.2043058	0	0.9603	88265	0.1905593	0.2086675	0	0.9603
literacy	63980	0.7045989	0.2402473	0.03	1	90503	0.7029143	0.2513853	0.03	1
logdist	63980	8.636535	0.8291331	1.899118	9.901581	90503	8.636368	0.8189654	1.899118	9.901581
comborder	63980	0.0265239	0.1606885	0	1	90503	0.0230711	0.1501301	0	1
comcur	63980	0.0133479	0.1147604	0	1	90503	0.0126405	0.1117176	0	1
rta	63980	0.022929	0.1496785	0	1	90503	0.0942179	0.2921332	0	1
comcol	63980	0.1011254	0.301497	0	1	90503	0.098063	0.2974015	0	1
col45	63980	0.0152235	0.1224418	0	1	90503	0.01074	0.1030764	0	1
colony	63980	0.0265083	0.1606425	0	1	90503	0.0184524	0.1345813	0	1
curcol	63980	0.0024852	0.0497897	0	1	90503	0.0003757	0.0193789	0	1
comleg	63980	0.3461394	0.475742	0	1	90503	0.3273483	0.4692481	0	1
relprox	60109	0.1739201	0.2249309	0	0.99	80498	0.1628158	0.2212375	0	0.99
war	60075	0.0111045	0.0854422	0	1.3	80496	0.0064885	0.0625127	0	1.3
loggdp	63869	35.08006	2.696889	24.51465	44.86049	90173	35.75253	2.747575	25.23617	46.33582
loggdppc	63869	17.11748	1.558097	11.57781	23.08225	90173	17.60787	1.785094	11.19864	22.8365
logpop	63869	17.96257	2.561552	8.234294	27.6137	90173	18.14465	2.55353	8.026937	28.10989
logarea	63980	23.9537	3.440518	7.358449	33.0341	90503	23.55652	3.378828	7.358449	32.77105
landlocked	63980	0.2164426	0.4407519	0	2	90503	0.3557009	0.54146	0	2

Table B.2: Correlations

1970-1991 (obs=48927)

	col	cnldummy	csldummy	cslnativedu	cslnonnati	cslnativenc	lp1	lp2	cnl	csl	cslnative	cslnonnati	cslnativenc	diversity	literacy	logdist	comborder	comcur	rta	comcol	col45	colony	curcol	comleg	relprox	war
col	1																									
cnldummy	0.5778	1																								
csldummy	0.4304	0.3283	1																							
cslnativedummy	0.5778	1	0.3283	1																						
cslnonnativedummy	-0.0157	-0.1327	0.5451	-0.1327	1																					
cslnativenonnativedummy	0.1378	-0.1398	0.4894	-0.1398	-0.1307	1																				
lp1	-0.1808	-0.0785	0.1045	-0.0785	-0.0437	0.3235	1																			
lp2	-0.2521	-0.1471	0.0343	-0.1471	-0.0588	0.2919	0.8639	1																		
cnl	0.5435	0.8133	0.267	0.8133	-0.1381	-0.1327	-0.1204	-0.1961	1																	
csl	0.6107	0.7197	0.5583	0.7197	0.0103	0.2517	0.0578	-0.0008	0.733	1																
cslnative	0.61	0.9178	0.3013	0.9178	-0.1488	-0.146	-0.1032	-0.1751	0.9266	0.7891	1															
cslnonnative	0.0192	-0.0701	0.3301	-0.0701	0.6057	-0.0494	0.0203	0.0114	-0.0699	0.2193	-0.0773	1														
cslnativenonnative	0.1314	-0.1084	0.3509	-0.1084	-0.0862	0.717	0.2574	0.2701	-0.0984	0.4422	-0.1087	-0.0421	1													
diversity	0.1407	-0.055	-0.0658	-0.055	0.0155	-0.0807	-0.2646	-0.2905	-0.1151	-0.1441	-0.1022	-0.0392	-0.0776	1												
literacy	-0.0762	0.0962	0.1552	0.0962	-0.0283	0.216	0.283	0.3712	0.1048	0.2384	0.1187	0.0881	0.1953	-0.4765	1											
logdist	-0.2053	-0.3032	-0.1964	-0.3032	-0.2121	0.101	-0.0416	0.0047	-0.2725	-0.2834	-0.2719	-0.243	0.0532	-0.1079	0.0944	1										
comborder	0.1375	0.2143	0.0899	0.2143	0.0448	-0.0182	0.0145	-0.019	0.1555	0.1309	0.1615	0.0354	-0.0381	0.0957	-0.067	-0.3609	1									
comcur	0.1777	0.0849	0.0937	0.0849	0.099	-0.0405	-0.0309	-0.059	0.0313	0.0507	0.0487	0.0711	-0.0253	0.1583	-0.118	-0.1969	0.0856	1								
rta	0.0724	0.1296	0.1225	0.1296	0.118	0.0627	0.1437	0.1385	0.09	0.2159	0.1098	0.2066	0.102	-0.0594	0.0971	-0.3163	0.0969	0.0935	1							
comcol	0.316	0.0726	0.1722	0.0726	0.1642	-0.0093	-0.098	-0.1189	0.0338	0.0578	0.0375	0.0835	-0.0015	0.2333	-0.1736	-0.1505	0.0633	0.2607	0.0961	1						
col45	0.176	0.0527	0.0775	0.0527	-0.036	0.1275	-0.0365	-0.038	0.0429	0.1136	0.0577	-0.0234	0.1281	-0.0473	0.002	-0.0056	-0.0113	0.0176	-0.0086	-0.0409	1					
colony	0.2072	0.1463	0.093	0.1463	-0.0341	0.0985	-0.029	-0.03	0.1421	0.1889	0.1618	-0.0273	0.102	-0.0674	0.0333	-0.0303	0.0462	0.0097	-0.0053	-0.0549	0.7758	1				
curcol	0.0806	0.0679	0.0427	0.0679	0.0043	0.0357	-0.0158	-0.0166	0.0577	0.0869	0.0644	0.0144	0.0477	-0.0248	0.0296	-0.0013	-0.0084	0.0394	-0.0082	-0.0177	0.3924	0.3045	1			
comleg	0.3853	0.2487	0.1916	0.2487	-0.0049	0.0656	-0.0536	-0.071	0.2563	0.2274	0.2808	-0.0176	-0.0223	0.1172	-0.1364	-0.1127	0.0969	0.1283	0.0445	0.293	0.1266	0.1518	0.0375	1		
relprox	0.2731	0.394	0.2083	0.394	-0.0558	0.0735	0.174	0.1324	0.4011	0.3509	0.4039	-0.0257	0.0115	-0.0365	0.0218	-0.2283	0.1825	0.0262	0.0749	0.0189	0.0078	0.0986	0.0144	0.2437	1	
war	-0.0413	0.0014	0.0769	0.0014	0.0923	0.0559	0.0319	0.0493	-0.0164	0.0312	-0.0158	0.0839	0.0346	-0.0616	0.1119	-0.1169	0.1167	-0.0163	0.0636	-0.0407	0.0006	0.0174	-0.0004	-0.0528	-0.009	1

1995-2011 (obs=71865)

	col	cnldummy	csldummy	cslnativedu	cslnonnati	cslnativenc	lp1	lp2	cnl	csl	cslnative	cslnonnati	cslnativenc	diversity	literacy	logdist	comborder	comcur	rta	comcol	col45	colony	curcol	comleg	relprox	war
col	1																									
cnldummy	0.5224	1																								
csldummy	0.4064	0.3449	1																							
cslnativedummy	0.5224	1	0.3449	1																						
cslnonnativedummy	0.0256	-0.1002	0.5806	-0.1002	1																					
cslnativenonnativedummy	0.1274	-0.1093	0.5161	-0.1093	-0.0799	1																				
lp1	-0.1691	-0.0594	0.1089	-0.0594	-0.0011	0.267	1																			
lp2	-0.2307	-0.0975	0.0588	-0.0975	-0.0083	0.2332	0.8403	1																		
cnl	0.5266	0.7462	0.2573	0.7462	-0.1023	-0.1002	-0.1023	-0.1606	1																	
csl	0.5694	0.6754	0.5823	0.6754	0.0807	0.297	0.0853	0.0408	0.7051	1																
cslnative	0.5835	0.8765	0.3023	0.8765	-0.1098	-0.1146	-0.0799	-0.1324	0.9006	0.7798	1															
cslnonnative	0.0381	-0.0557	0.3617	-0.0557	0.6231	-0.0299	0.0523	0.0592	-0.0521	0.2589	-0.061	1														
cslnativenonnative	0.1129	-0.092	0.3804	-0.092	-0.0511	0.737	0.2486	0.2444	-0.0774	0.4638	-0.0898	-0.0247	1													
diversity	0.1274	-0.0286	-0.0842	-0.0286	-0.0282	-0.0885	-0.2825	-0.3034	-0.0977	-0.1441	-0.0812	-0.0647	-0.099	1												
literacy	-0.11	0.0913	0.1639	0.0913	0.0087	0.2063	0.303	0.3921	0.07	0.2121	0.0926	0.1004	0.1843	-0.4634	1											
logdist	-0.134	-0.2727	-0.2007	-0.2727	-0.2327	0.0717	-0.0936	-0.1108	-0.2183	-0.2516	-0.2275	-0.2728	0.0546	-0.0833	0.0072	1										
comborder	0.1156	0.2157	0.1092	0.2157	0.0648	0.0021	0.0412	0.0363	0.1463	0.1544	0.1657	0.0606	-0.0108	0.0739	-0.0261	-0.363	1									
comcur	0.1425	0.0474	0.1137	0.0474	0.158	0.0055	0.0068	-0.0212	0.0158	0.0706	0.0256	0.135	0.0092	0.1382	-0.0884	-0.2072	0.1294	1								
rta	0.0561	0.1352	0.2354	0.1352	0.2243	0.0964	0.1246	0.1288	0.0867	0.2289	0.106	0.2817	0.0867	-0.0465	0.1343	-0.4462	0.2134	0.1845	1							
comcol	0.323	0.1419	0.1971	0.1419	0.1179	0.035	-0.0737	-0.078	0.0495	0.1085	0.094	0.0619	0.0159	0.2103	-0.1381	-0.1395	0.0814	0.2039	0.0864	1						
col45	0.158	0.0913	0.0809	0.0913	-0.0205	0.0904	-0.0238	-0.0259	0.0704	0.1366	0.1029	-0.0177	0.1003	-0.044	0.017	-0.0162	0.0386	-0.0121	0.032	-0.0312	1					
colony	0.1684	0.1465	0.1093	0.1465	0.0021	0.0852	-0.0198	-0.0191	0.1361	0.1898	0.1652	-0.0041	0.093	-0.0606	0.0408	-0.0462	0.0879	-0.0164	0.0461	-0.0446	0.7372	1				
curcol	0.0294	0.0545	0.0205	0.0545	0.0178	-0.0053	-0.0107	-0.0096	0.0641	0.051	0.0499	0.0316	-0.0028	-0.0133	0.0159	-0.0254	-0.003	-0.0022	0.0127	-0.0063	0.1814	0.1337	1			
comleg	0.3731	0.2515	0.1982	0.2515	0.0166	0.0745	-0.0408	-0.0248	0.2333	0.2271	0.2709	0.0025	-0.0066	0.0921	-0.1386	-0.1064	0.1082	0.1116	0.0505	0.338	0.1185	0.1146	0.002	1		
relprox	0.2503	0.3259	0.1734	0.3259	-0.0274	0.0592	0.1363	0.1025	0.3507	0.3064	0.3416	-0.0025	0.0316	-0.0067	-0.0288	-0.1927	0.1495	0.0573	0.0961	0.043	0.0097	0.0788	0.0203	0.1983	1	
war	-0.0304	0.0057	0.0891	0.0057	0.101	0.0792	0.0581	0.0634	-0.0092	0.054	-0.0104	0.0924	0.0599	-0.0519	0.0905	-0.108	0.1203	0.0865	0.1395	-0.0316	0.0005	0.0258	0.0071	-0.0355	-0.0028	1