Language barriers to foreign trade: evidence from translation costs*

Alejandro Molnar†
Vanderbilt University
DRAFT, NOT FOR CIRCULATION

November 9, 2013

Abstract

[Foreign trade involves tasks that may be subject to language barriers, such as researching foreign markets, communicating with counterparties and marketing products to foreign consumers. Language skills have a wage premium that is determined by local and worldwide supply and demand for language services, and the premium is specific to each country and pair of languages. I construct a novel measure of language skill premia based on professional rates for translation services from an online market. The skill premium measure relies on the bi-directional nature of the translation cost data to control both for difficulties inherent in defining a unit of account (as the per word "piece-rates" common in the translation industry do not embody equal amounts of work across languages) and the skilled-wage component of rates. I develop an estimation strategy based on overlaps in ethnolinguisitic populations to estimate the effect of the language skill premium as a cost barrier to trade, net of confounders such as trade by shared ethnic populations. I find that accounting for country and language-specific language barriers yields a three-fold increase in the estimated effect of language on foreign trade, relative to current estimates based on a shared common language.]

JEL: F1, F14, F23, Z13

Keywords: Gravity equation, common language, language barriers.

*I am grateful to Tim Bresnahan, Jon Levin and Kalina Manova for guidance and encouragement. Special thanks to Ran Abramitzky, Kyle Bagwell, Emilio Depetris Chauvin, Bernardo Díaz de Astarloa, Doireann Fitzgerald, Gordon Hanson, Han Hong, Paul Ma and Andrés Rodríguez-Clare for helpful comments and discussions. All remaining errors are my own.

†Department of Economics, Vanderbilt University, 415 Calhoun, Nashville, TN 37204 , e-mail: alejandro.i.molnar@vanderbilt.edu.

1 Introduction

In this paper I study the relationship between the languages that are spoken in a country and the country's patterns of foreign trade. Foreign trade involves tasks that may be subject to language barriers, such as researching foreign markets, communicating with counterparties and marketing products to foreign consumers. Language skills have a wage premium that is determined by local and worldwide supply and demand for language services, and the premium is specific to each country and pair of languages. The main idea in this paper is to exploit variation in country-specific prices for translation services across different pairs of languages to recover country-specific measures of language skill premia. I construct a novel measure of these premia based on professional rates for translation services from an online market. The skill premium measure relies on the bi-directional nature of the translation cost data to control both for difficulties inherent in defining a unit of account (as the per word "piece-rates" common in the translation industry do not embody equal amounts of work across languages) and the skilled-wage component of rates. I develop an estimation strategy based on overlaps in ethnolinguisitic populations to estimate the effect of the language skill premium as a cost barrier to trade, net of confounders such as trade by shared ethnic populations. I find that accounting for country and language-specific language barriers yields a three-fold increase in the estimated effect of language on foreign trade, relative to current estimates based on a shared common language.

2 Online markets for translation services

In this section I describe the offline and online markets for translation services, as well as the price data available from online translation markets and how it contains information on skill premia for a specific form of human capital that is used intensively in foreign trade.

The task of translation is typically carried out by a single individual. A translator's physical productivity can be thought of as the pace at which a source text is translated into a target text of a given quality. A translator's productivity is text-dependent, as translation work may require domain-specific knowledge in addition to language skills, and text may vary in difficulty, requiring additional time (e.g. for research) to produce a translation of a given quality.

The typical translator is a freelance worker, but relationships with the demand side of the translation market may be in-house as well as arms length. Relationships may

be long- or short-term, and occur directly or through intermediaries called translation agencies. Agencies provide demand risk-sharing among teams of individual translators, as well as reputation services (on reputation intermediaries in online labor markets, see Stanton and Thomas, 2011).

The price of translation work is called a "translation rate" and is quoted in a unit of text that can be easily counted on a computer: words in "spacing" languages (i.e. those in which words are separated by white space such as English) and characters in non-spacing languages (e.g. Simplified Chinese). Translation rates are specific to a directed language pair (e.g. Spanish to English) and are almost always expressed in units of the source text, as quoting per unit in the target language provides bad incentives for the translator. As freelancers, translators negotiate rates bilaterally with potential clients and will usually quote client-specific rates based on the current state of demand for their services and the attributes of the client: for example, translators may quote higher rates for rushed jobs or highly technical jobs, or lower rates when providing quantity discounts or attempting to establish a relationship with a client that may be a future source of demand. Offline demand for a translator can come from professional listing services, translation agencies, reputation and word of mouth. A further source of demand can be outsourcing from other translators, and these translators provide editing and monitoring and may or may not disclose to the final client that the translation was outsourced. Most demand from these offline sources is specific to a translator's country of residence.

Online platforms specifically designed to intermediate global translation markets started in 1999 (proz.com) and 2002 (translatorscafe.com). These platforms introduced new forms of market organization (e.g. procurement auctions, explicit reputation metrics for both sides of the market) and greatly facilitated the meeting of supply and demand across borders. To do business on these platforms, translators must create a profile and report the minimum translation rate at which they are willing to work on each language pair in which they work. These rates are not revealed to potential clients, but screen the jobs that a translator will see on a job listings dashboard when logged into the site. Since this screening rate is set prior to and independently of the attributes of listed translation jobs, it can be thought of as a reservation wage (in units of work, rather than time) for each translator. The second largest of these online markets, translatorscafe.com, discloses the average of this minimum translation rate for all translators in a language pair that are located in a particular country.

There is substantial variation in these average rates both across language pairs within country and across countries within language pair. The determinants supply and

demand for translation services that give rise to variation across language pairs within country arises are reasonably straightforward: for example, demand may depend on the languages used by trade partners, foreign tourists, and on the country's interest in cultural products produced in foreign languages, and these in turn may depend on the country's ethnolinguistic composition. Supply follows from each country's endowment of people with language skills that enable work in each specific language pair (which may be relatively fixed in the short term), and an opportunity cost of time for this type of work, common across all possible translation pairs (e.g. the wage for the bundle of skills that translators possess net of their language-specific skills). Variation across countries in the same language pair is less straightforward, as a law of one price might be expected to hold in online markets. One reason for cross country variation in prices is that translation services are not homogeneous and may require country-specific knowledge (e.g. on the legal environment, pop culture, slang or vernacular) for which translators in different countries are imperfect substitutes. A second reason follows from the microstructure of the translation industry and the fact that translators have limited capacity and spend a fraction of their time unemployed and waiting for the arrival of the next job. Accepting a low-paying job removes the translator from the market until the job is completed, and therefore may preclude accepting a job that arrives later with a higher pay. Online translators face such arrival processes for potential jobs from both online and offline sources, and a simple search model suggests that they should set a reservation wage for accepting an online job that depends on the opportunity cost of removing themselves from the offline market for a period of time.

A further idiosyncratic factor of this industry is that the unit of work in which prices are expressed is not constant across languages. Different western languages may differ in their use of articles, contractions and compound words, so that texts that are supposed to convey the same "meaning" would be counted at different lengths depending on the language in which they are written. For instance, the English phrase "it is not so simple" consists of 5 words, whereas the same phrase in Spanish "no es tan sencillo" consists of 4.[1] This applies even more clearly to rate comparisons between spacing and non-spacing languages, where rates are not expressed in words.

[1]The phrase in English can be contracted to "it's not so simple" or "it isn't so simple", both of which are also 4 words long. On average, text in Spanish tends to be longer than equivalent text in English. German uses long compound words, and a famous example that arose from a state legislature was the "Rindeischetikettierungsueberwachungsaufgabenuebertragungsgesetz", or the "Law on delegation of duties for supervision of beef labeling". If the amount of work required from a translation is not a function of the number of words but of the amount of "meaning" conveyed in the text, we should expect, all else equal, that translation rates out of German be higher than out of English, and these higher than out of Spanish.

To develop a comparable measure across languages, I assume the following structure for observed translation rates, which are quoted in US dollar cents per source word or character:

$$r_{abc} = \delta_a \delta_{(a,b)c} \eta_{abc} \tag{1}$$

where r_{abc} is the observed average translation rate from source language a to target language b for translators located in country c, δ_a is a scaling factor specific to the source language intended to absorb differences between languages in the amount of work embodied in translating a word or character, and $\delta_{(a,b)c}$ is an undirected pair and country effect. I normalize $\log \delta_{(eng,spa)USA} = 0$ and regress $\log r_{abc}$ on source language fixed effects and undirected pair and country fixed effects, so average rates by source language relative to the rate on this specific pair are absorbed by source effects, and remaining variation in rates is captured by the $\delta_{(a,b)c}$ term. For example, the value of $\delta_{(eng,spa)ARG}$ is constructed from the average of rates for English to Spanish (net of the English source language effect) and Spanish to English (net of the Spanish source language effect) for translators located in Argentina.

After adjusting for source effects, a country's average translation rate on a language pair remains a nominal quantity. Equal nominal rates can represent widely different resource costs for countries with different average wage levels for skilled labor. For example, the average English to French rate is 10 USD cents per word for translators located in France, 13 cents in Côte d'Ivoire, 12 cents in Algeria, 10 cents in Morocco, 8 cents in Cameroon and 7 cents in Senegal. As France's GDP per capita is approximately 40 times that of Senegal, the resource cost of employing a person to overcome English to French language barriers is presumably substantially higher as a per-person share of Senegal's economy than it is for the French economy, and much more so for the economy of Cote d'Ivoire.[2]

To obtain a measure of variation across language pairs that is net of wages, I regress $\widehat{\log \delta_{(a,b)c}}$ on GDP per capita in country c, and take the residuals of this regression as my final measure of real language skill premia. Figure **??** plots this exercise, plotting $\widehat{\log \delta_{(a,b)c}}$ in blue for language pairs that involve a country's most widely spoken language, and green otherwise. Some of the highest language skill premia include English-French in Cote d'Ivoire, English-Swahili in Tanzania, and Italian-Japanese in

[2]Similarly, translation rates between Scandinavian languages are relatively high, presumably because most speakers of these languages reside in high wage countries. A reasonable prior is that language barriers between Scandinavian counties are relatively low.

Japan. Some of the lowest include English-Farsi in Afghanistan, English-Khmer in Cambodia, Kazakh-Russian in Kazakhstan and English-Albanian in Greece. Figure 2 plots a subset of this data, narrowing in on countries for which English is not the most widely spoken language and the skill premium for English and the country's most widely spoken language.

2.1 Why language could matter in foreign trade

Trade requires communication, which can give rise to language barriers. Language-intensive tasks that are essential to trade include researching foreign markets, adapting and marketing products to foreign consumers,[3] and communication and contracting between importing and exporting firms. Language may affect the foreign direct investment decisions of multinational firms, e.g. the location of regional headquarters, and trade in final and intermediate goods may follow from such decisions.

Empirical estimates of trade costs acknowledge the role of language (**?**), usually estimated by inclusion of a binary variable for whether two countries share an official language. The size of language barriers can be expected to depend on the language-intensity of specific tasks involved in trade, and the cost of hiring workers within a country with the required language skills.

3 Empirical evidence on language barriers to trade

In this section I describe how my measure of language skill premia explains trade flows between countries by including the language skill premia described in Section 2 in standard trade gravity estimation frameworks from the trade literature. I also describe an empirical strategy to estimate a causal effect of language barriers on trade.

In order to include the language and country-specific adjusted translation rates in a standard gravity equation framework, I map the adjusted rates to country pairs in two ways. I define the *top pair* country-pair specific rate as the average of the rates between the most widely spoken languages in each country. I also assign a value of 1 to

[3]Firms make design or product choices in response to language barriers, an example of which are the text-free assembly manuals that accompany furniture sold across many national markets by Swedish firm IKEA. To sidestep individual language costs, IKEA incurs the cost of high quality design in assembly manuals (and perhaps product adaptation) to avoid ambiguity and the use of written text. Not all retail products that require instruction manuals or assembly instructions in a language other than that of design or manufacture are suitable to, or have sufficient scale to afford, IKEA's economy of words. See Kelly and Zetzsche (2012) for examples of how translation services are employed in trade.

the *translation rate observed* dummy for the pair.[4] I construct a *fractional* or population-weighted adjusted translation rate by applying the above procedure to every language population within a country combined with every language population in a partner country. The translation rate measure is obtained from the weighted average of every cell with a combination of languages, where the weight is the product of the marginal population measures in each country. Since rates for most potential language pairs are not observed, the measure of population for which the rate is observed is counted in the continuous *translation rate observed* variable.[5]

Additional linguistic variables include the measure of the population in a country pair that speak the same language, i.e. the probability that a person picked at random from one country would be able to speak with a person picked at random from the other. This is equivalent to the main measure used in the preceding work on the effect of language on trade by Melitz (2008) and Melitz and Toubal (2012). I include the log of this variable and a dummy for when countries share no speakers of a common language. I also include the dummy variable for *common official language*, which is the usual control for language used in almost all prior empirical work that estimates the gravity equation.

Table 1 reports OLS estimates for three specifications of the standard gravity equations for the subset of country pairs with positive trade flows. The first column includes only the standard *common language* dummy, the second includes the fractional adjusted rate measures and the third only the rates for the top pair of languages.[6] From the estimates in Column (2), a 1 percent increase in the adjusted translation rate between the languages of a pair of countries is associated with a 0.6 percent decline in bilateral trade. Column (3) presents a weaker, non-significant estimate of the same effect for the most widely spoken language for each country in the pair. In both cases the magnitude of estimates for distance, contiguity and colonial relationships all decline substantially after inclusion of the larger set of language skill premium and linguistic overlap measures. The magnitude for "common official language" is almost halved, but inclusion of the measures of ethnolinguistic overlap means that this binary variable identifies purely the "official" status of any common languages.

Table 2 presents results from an exponential regression on the same data, which allows inclusion of pairs with zero trade that are dropped due to the log transformation, and is a favored empirical method in the trade literature because gravity equation

[4]If language pair data for only one country is observed, I include the rate and count it as observed.

[5]Some translation rates in the data that are plotted in Figure **??** do not map to any ethnic population (e.g. English to Spanish in Norway) and are therefore not used in gravity equation estimates.

[6]Results are robust to netting GDP per capita from nominal rates with a quadratic term or a local polynomial regression.

regressions are motivated by multiplicative structural models and estimates from the exponential regression framework are robust to heteroscedasticity in the error term for the parameters of interest in these models. Both this framework and the framework I will employ below to instrument for the language measures cannot include importer and exporter fixed effects, so I include instead a "remoteness" measure (used for example by Baldwin and Harrigan (2011) and Manova and Zhang (2012), see discussion in Head and Mayer (2013)). The main coefficient estimates on this sample are similar in magnitude, but the estimates from the Column (3) specification are now significant. Common official language becomes non-significant, and the inclusion of the full set of language variables has a smaller effect on the coefficients for geographical covariates such as distance and contiguity.

The language barrier estimates from Tables 1 and 2 cannot be given the causal interpretation of a trade cost because linguistic overlap can be correlated with trade through other channels, in particular ethnic trade (Rauch and Trindade, 2002). If trade creates additional demand for translation services on the relevant language pairs and this increases translation rates, this would lead trade to be positively correlated with observed translation rates, biasing upwards (in this case, towards attenuation) the coefficient estimates for translation rates as a trade cost in a gravity equation.

I use data on the overlaps of ethnolinguistic populations to develop a shifter for the relative scarcity of language skills and estimate a causal effect of country-specific language skills on trade. To describe the empirical strategy, consider the example of trade between the United Kingdom and either Vietnam or Thailand. The population of Vietnamese and Thai speakers in the UK is small and similar in magnitude, but there is one large difference between these language pairs from the perspective of the UK market for language services: the existence of the US as a majority English-speaking country with a large population of Vietnamese speakers. There is no similar example for the Thai language. The existence of the US as a country where Vietnamese and English speakers overlap has an effect on the world market for language services in this language pair. In particular, if the ethnolinguistic composition of the US has a larger effect on supply than on demand for English-Vietnamese language services, this is likely to decrease the English-Vietnamese language skill premium in the UK.

Based on this idea, I define an ethnic overlap instrument in the following manner: for a given "language in country" pair (e.g. English in the UK, Vietnamese in Vietnam), the probability that a speaker of each of these languages would coexist in an country other than that of the language pair with a speaker of the other language. That is, I take the measure of all English speakers worldwide with the exception of the UK, and

calculate the conditional probability that if they were to meet a fellow resident at random from their country, that resident would be a speaker of Vietnamese. This conditional probability is close to zero for most English speakers worldwide, except for the US where it is about 0.02, and the US constitutes a large fraction of worldwide English-speakers. I then do the same for Vietnam, where the measure is almost zero in neighboring countries with Vietnamese speaking residents, but a large fraction of Vietnamese speakers outside the UK-Vietnam pair reside in the US, where their probability of meeting an English speaker is close to 1. I construct the instrument as the product of both of these probabilities.

As this instrument coarsens the variation in the data to a language-pair level (since foreign overlap will be a similar value for most countries, e.g. the instrument has a similar value for English-Vietnamese pairs that involve Vietnam and Great Britain, Australia or New Zealand), the estimated effect might be expected to be similar to that from an OLS regression of trade flows on a language pair (but not country) specific measure. However, whereas the average global rate may be driven by trade flows on a particular country (e.g. the abundance of Vietnamese speakers in the US affects both the worldwide supply and demand for English-Vietnamese language services, as well as trade between the US and Vietnam through a direct ethnic channel, see Rauch and Trindade, 2002), the instrument specifically excludes the correlation between trade flows and a country pair's common language fraction (e.g., in the case of US to Vietnam trade, the instrument takes on a low value, whereas the common language fraction takes on a relatively high value).

The fact that I do not observe language skill premia in all language pairs presents a hurdle to straightforward instrumental variables estimation, as the common practices of selecting a sample or interacting unobserved or censored values of a regressor with a dummy variable (as I do in the results presented in Tables 1 and 2) are only valid under exogeneity assumptions. When a censored regressor is endogenous, 2SLS estimation can lead to a bias that amplifies the magnitudes of estimated coefficients, as discussed by Rigobon and Stoker (2007).[7]

As detailed in Appendix Appendix B, I implement the estimation strategy of Chernozhukov, Rigobon, and Stoker (2010) to estimate the effect of language skill premia on trade using the ethnolinguistic overlap instrument. Results are in Table 3, both for the "conditional on positive" linear regression specification and the exponential regression. The bottom row in the table presents results for the first stage: for the language

[7]In fact, estimating Columns 2 and 3 of Table 1 by 2SLS with the ethnolinguistic overlap instrument does lead to implausibly large effects of language skill premia on trade flows

pair of a specific country, ethnolinguistic overlap elsewhere in the world predicts a lower median adjusted translation rate, which suggest that the effects of ethnolinguistic overlap through the expansion of supply dominates the effects through expansion of demand. Any effect from the ethnolinguistic overlap for the common pair in question is absorbed by the *common language fraction* variable, which is included in the second stage. The main coefficient estimates for the effect of language skill premia on trade are larger in magnitude than in the specifications of Tables 1 and 2, which is consistent with the main concern for endogeneity in those regressions being attenuation from the reverse the effect of trade on translation costs. The magnitudes from the exponential regression specification imply that predicted trade for a pair of countries sharing a common language is 5.2 times that of a pair of countries not sharing a common language and having the largest language skill premium observed in the data, relative to a 1.7-times increase estimated using only the common language dummy. This is a three-fold increase in the estimated effect of language on foreign trade. Results are robust across specification that use alternative thresholds for translator data quality and when I include data from a second source (translationdirectory.com, an online listing directory for translators, where incentives to price revelation are not as clear).

[**To be done**: Classification of products (e.g. more differentiated products, contract-intense, R&D intense, etc.). Preliminary results show monotonic results between product differentiation or technological component and impact of measure. Pairs through English as a *lingua franca*. Extensive vs. intensive margins with WB Exporter Dynamics data. Alternative instruments based on migration flows and linguistic cleavages (e.g. Shastry, 2012). List of potential alternative channels, e.g. Warcziarg-style covariates, and Guiso, Sapienza, and Zingales (2009). Recast specification into the ideal for comparability of OLS and control function results, which is to "splinter" the gravity equation into population-proportional subcells to use the population-weighted language barrier measure between country pairs. For OLS show equivalence of splinter-regression to standard language-population-weighted]

4 Conclusions

This paper develops a novel measure of language barriers between countries based on prices for translation services, which reflect the market premium on scarce language skills, and estimates the impact of this measure on trade flows between countries. The paper's main result is that the conventional practice of controlling for language with a "common official language" dummy omits a large share of the effect of language

on foreign trade. Understanding the proper role of language contributes to our understanding of the barriers to trade and economic integration that some countries may face: much as 'landlocked" countries trade less, firms in countries with low endowments of foreign language skills may face additional hurdles to carrying out the multiple activities involved in foreign trade, and its domestic firms may need to rely on the initiative of foreign partners to overcome these barriers. Additional knowledge on the component factors that are regularly proxied by "distance" in standard gravity applications may reduce the relevance attributed to this catch-all variable, and increase our understanding of the nuanced factors that affect foreign trade. As an institutional and cultural endowment, the abundance of language skills may have broader implications for other flows such as the transmission of technological knowledge and cultural values.

References

Baldwin, R., and J. Harrigan (2011): "Zeros, Quality, and Space: Trade Theory and Trade Evidence," *American Economic Journal: Microeconomics*, 3(2), 60–88.

Chernozhukov, V., and H. Hong (2002): "Three-Step Censored Quantile Regression and Extramarital Affairs," *Journal of the American Statistical Association*, 97(459), 872–882.

Chernozhukov, V., R. Rigobon, and T. M. Stoker (2010): "Set identification and sensitivity analysis with Tobin regressors," *Quantitative Economics*, 1(2), 255–277.

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

Hallak, J. C. (2006): "Product quality and the direction of trade," *Journal of International Economics*, 68(1), 238–265.

Head, K., and T. Mayer (2013): "Gravity Equations: Workhorse, Toolkit, and Cookbook," *CEPR Discussion Paper No. 9322*.

Kelly, N., and J. Zetzsche (2012): *Found in Translation*. Perigee Books.

Kugler, M., and E. Verhoogen (2012): "Prices, Plant Size, and Product Quality," *The Review of Economic Studies*, 79(1), 307–339.

Manova, K., and Z. Zhang (2012): "Export Prices Across Firms and Destinations," *The Quarterly Journal of Economics*, 127(1), 379–436.

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

Melitz, J., and F. Toubal (2012): "Native language, spoken language, translation and trade," .

Rauch, J. E. (1999): "Networks versus markets in international trade," *Journal of International Economics*, 48(1), 7–35.

Rauch, J. E., and V. Trindade (2002): "Ethnic Chinese networks in international trade," *Review of Economics and Statistics*, 84(1), 116–130.

Rigobon, R., and T. M. Stoker (2007): "Estimation with Censored Regressors: Basic Issues*," *International Economic Review*, 48(4), 1441–1467.

Shastry, G. K. (2012): "Human Capital Response to Globalization Education and Information Technology in India," *Journal of Human Resources*, 47(2), 287–330.

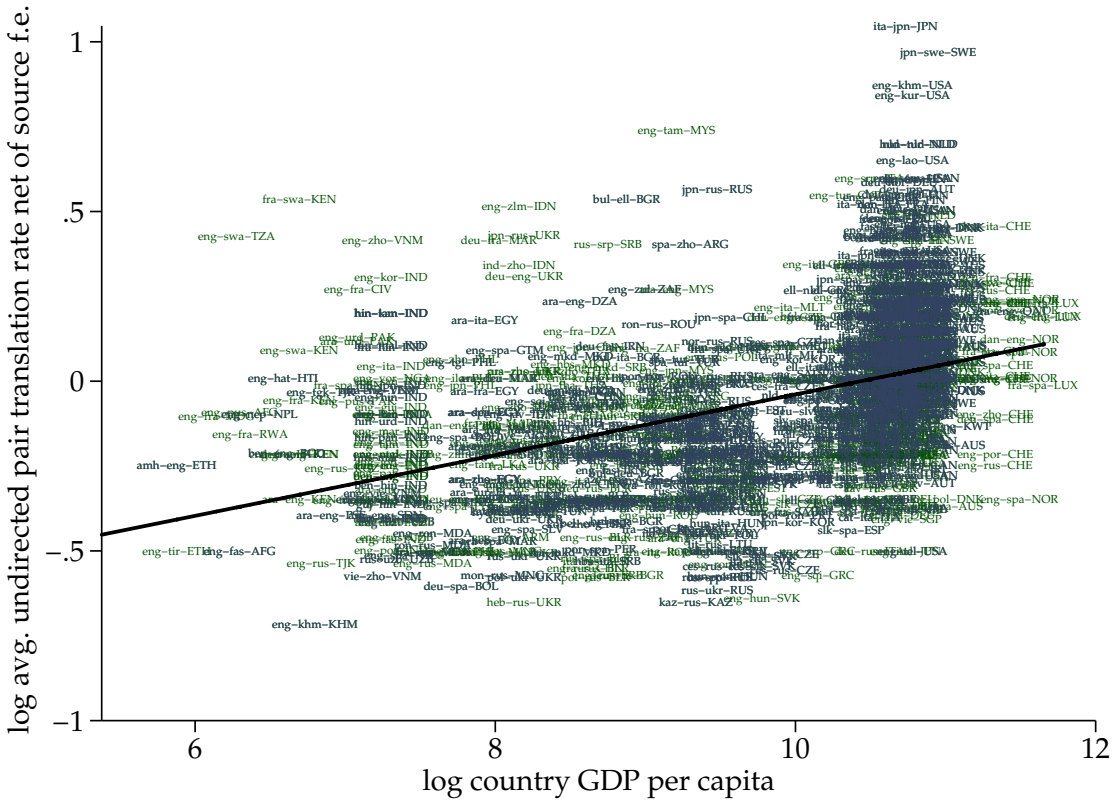
Silva, J. S., and S. Tenreyro (2006): "The log of gravity," *The Review of Economics and Statistics*, 88(4), 641–658.

Stanton, C., and C. Thomas (2011): "Landing the First Job: The Value of Intermediaries in Online Hiring," SSRN Scholarly Paper ID 1862109, Social Science Research Network, Rochester, NY.

Tables and figures

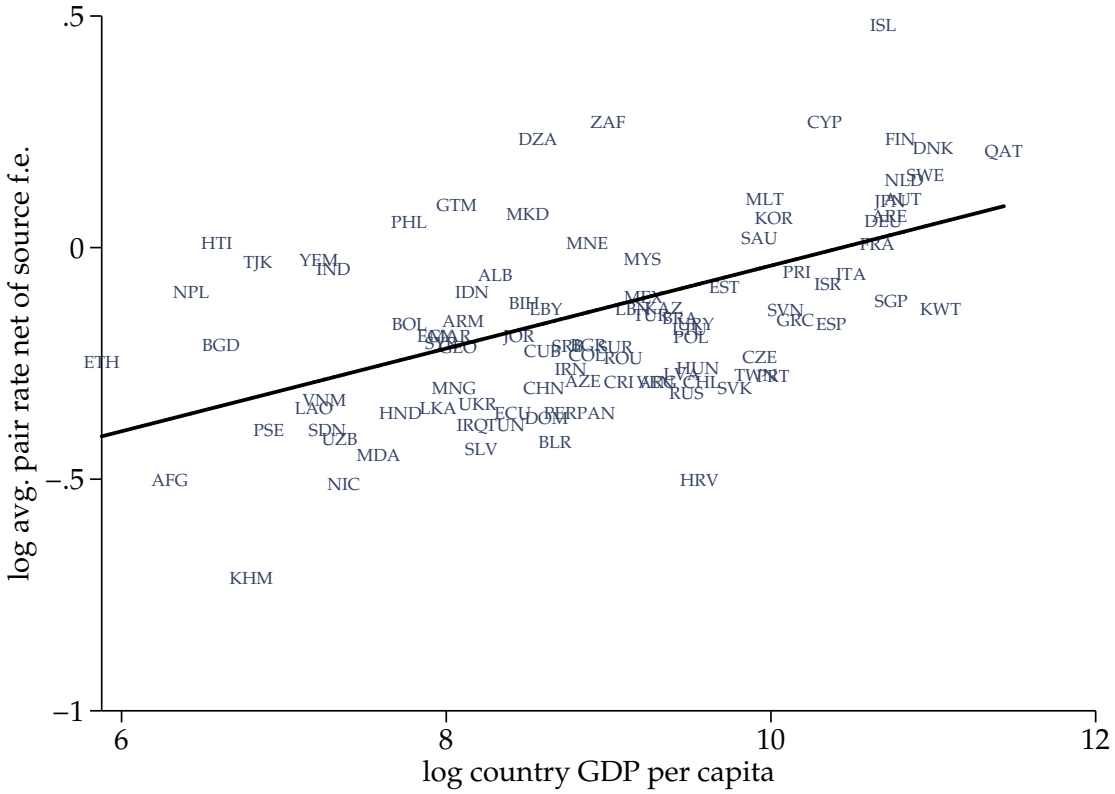
[To do: table of main languages, language iso codes, number of translators, average rate against English, Spanish, French, Chinese, English in the US.]

Figure 1: Adjusted translation cost and per capita GDP



Vertical axis: log of the average translation rates between a pair of languages (in both directions), net of source language fixed effects). English-Spanish rates for translators located in the United States are normalized to zero. Sample is coded by color: language pairs that involve a country's most widely spoken language in blue, language pairs that do not involve a country's most widely spoken language in green. Black line is linear fit.

Figure 2: Adjusted translation cost to English and per capita GDP



Vertical axis: log of the average translation rates between a country's most widely spoken language and English (undirected rates, i.e. combining rates where English is the source or the target language), net of source language fixed effects, plotted for countries for which English is not the most widely spoken language. English-Spanish rates for translators located in the United States are normalized to zero. Black line is linear fit.

Table 1: Trade and language barriers. Gravity linear regression on positive trade flows

	(1) Log trade$_{do} > 0$	(2) Log trade$_{do} > 0$	(3) Log trade$_{do} > 0$
Adj. translation rate (fractional)		-0.603*	
		[0.301]	
Translation rate observed (fractional)		0.370***	
		[0.075]	
Adj. translation rate (top pair)			-0.215
			[0.242]
Translation rate observed (top pair)			0.362***
			[0.060]
Log fraction common language		0.054***	0.055***
		[0.005]	[0.005]
No common language		-0.936***	-0.938***
		[0.064]	[0.064]
Common official language	0.587***	0.370***	0.376***
	[0.049]	[0.055]	[0.055]
Log distance	-1.517***	-1.401***	-1.403***
	[0.021]	[0.022]	[0.022]
Contiguity	0.809***	0.605***	0.605***
	[0.102]	[0.100]	[0.100]
Colonial tie (ever)	0.266*	0.065	0.090
	[0.128]	[0.130]	[0.130]
Colonial tie (after 1945)	1.209***	1.254***	1.223***
	[0.169]	[0.169]	[0.169]
Common colonizer (after 1945)	0.726***	0.614***	0.609***
	[0.064]	[0.065]	[0.065]
Exporter and importer f.e.	Yes	Yes	Yes
Observations	23767	23767	23767
R^2	0.95	0.95	0.95

Standard errors in brackets
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Trade and language barriers. Gravity exponential regression on all trade flows

	(1) Trade_{do}	(2) Trade_{do}	(3) Trade_{do}
Adj. translation rate (fractional)		-0.609*	
		[0.266]	
Translation rate observed (fractional)		-0.213*	
		[0.094]	
Adj. translation rate (top pair)			-0.506*
			[0.235]
Translation rate observed (top pair)			-0.095
			[0.073]
Log fraction common language		0.010	0.008
		[0.008]	[0.008]
No common language		-0.306*	-0.284*
		[0.128]	[0.126]
Common official language	-0.080	-0.184	-0.168
	[0.109]	[0.110]	[0.107]
Log distance	-0.685***	-0.677***	-0.663***
	[0.038]	[0.039]	[0.038]
Contiguity	0.566***	0.548***	0.550***
	[0.115]	[0.108]	[0.110]
Colonial tie (ever)	-0.147	-0.231	-0.213
	[0.135]	[0.144]	[0.143]
Colonial tie (after 1945)	0.263	0.303	0.312
	[0.233]	[0.262]	[0.258]
Common colonizer (after 1945)	0.532**	0.508**	0.541***
	[0.163]	[0.162]	[0.162]
Log gdp and remoteness (o & d)	Yes	Yes	Yes
Observations	41412	41412	41412
Log lik.	-8.45e+09	-8.36e+09	-8.38e+09

Standard errors in brackets

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Two-stage estimates of language barrier effect

	(1) Log $\text{trade}_{do} > 0$	(2) Trade_{do}
Second stage:		
Adj. translation rate (top pair)	-0.670*	-1.047***
	[0.263]	[0.283]
Log fraction common language	-0.004	0.016
	[0.012]	[0.011]
No common language	-0.505**	-0.389*
	[0.157]	[0.155]
Common official language	0.409*	-0.068
	[0.180]	[0.192]
Log distance	-1.162***	-0.644***
	[0.068]	[0.044]
Contiguity	0.950***	0.790***
	[0.196]	[0.108]
Colonial tie (ever)	0.243	-0.168
	[0.276]	[0.178]
Colonial tie (after 1945)	0.864**	0.245
	[0.288]	[0.282]
Common colonizer (after 1945)	0.046	0.595
	[0.410]	[0.406]
Log gdp and remoteness (o & d)	Yes	Yes
First stage:		
Log ethnolinguistic overlap	-.066	-0.062
	(0.021)	(0.021)
Observations	3554	4418
R^2	0.71	0.71

Standard errors in parentheses for coefficient on adjusted translation rate (top pair) is block bootstrapped at the country-pair level, and for first stage coefficient on log ethnolinguistic overlap is bootstrapped. Conventional standard errors in brackets are preliminary.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix A Data sources

Translation rates: Baseline results are from average rates by language pair and country, downloaded from translatorscafe.com on March 6, 2013 and discussed in the main text. Rates are included and labeled as "observed" if there are at least five translators present in a directed language-pair and country.

Trade flows: Trade flows for 2011 at the HS6 level are from the BACI dataset provided by CEPII, which are in turn based on United Nations Comtrade data.

Gravity covariates: Distance, common official language and colonial link data are from the CEPII Gravity Dataset.

Ethnolinguistic data: Counts of population by language group within country are from *Ethnologue, 16th Edition*.

R & D and advertising intensity: From Kugler and Verhoogen (2012), the ratio of advertising plus research and development expenditures to total sales, from the U.S. Federal Trade Commission (FTC) 1975 Line of Business Survey. Measures for ISIC 4-digit rev. 2 classification concorded to HS6.

Measure of horizontal differentiation: Classification due to Rauch (1999). SITC 4-digit industries concorded to HS6.

Appendix B Estimation with a censored endogenous regressor

I follow the method of Chernozhukov, Rigobon, and Stoker (2010) for estimation of a linear conditional mean model with a bound-censored and endogenous regressor. To describe the estimation approach, assume

$$\ln X = \beta L^{*} + D'\delta + U^{*} \tag{2}$$

$$L^{*} = Z'\pi + V^{*} \tag{3}$$

$$U^{*} = \gamma V^{*} + \varepsilon \tag{4}$$

where ε is mean independent of (V^{*}, L^{*}) and V^{*} is median independent of Z. The dependent variable X stands for exports and L^{*} is the uncensored language skill premium, which is endogenous when $\gamma \neq 0$. D is a vector of standard gravity regressors such as distance, and Z is a vector of instruments that includes D. We do not observe L^{*} for all pairs of languages, so for all unobserved pairs I set the language skill premium at its highest observed value $\overline{L}$ and assume that this is an upper censoring threshold such that an observed, censored language skill premium L is given by

$$L = \begin{cases} L^* & \text{if } L^* < \overline{L} \\ \overline{L} & \text{otherwise} \end{cases} \tag{5}$$

I estimate equation (3) in a first stage by censored quantile regression, employing the method of Chernozhukov and Hong (2002). Residuals from this first stage can be used as a control function for inclusion in a second stage, which can be estimated on the subsample above the censoring threshold. Construction of the control function for a linear conditional mean model follows directly. Applying the control function approach to the exponential conditional mean model for the gravity equation (as in Silva and Tenreyro, 2006) requires additional assumptions. I modify equation (2) to

$$\mathbb{E}\left[X|L^*, D, V^*\right] = \exp\left(\beta L^* + D'\delta + \gamma V^*\right). \tag{6}$$

Inclusion of a control function γV^* in equation (2) is a stronger functional form assumption, for which a sufficient condition is joint normality of (U^*, V^*). As censored quantile regressions are difficult to estimate with fixed effects, I replace exporter and importer fixed effects with importer and exporter gross domestic products and "remoteness" measures, following Baldwin and Harrigan (2011). Alternative gravity estimation methods aimed at removing country fixed effects (e.g. tetrads, see Hallak (2006) and Head and Mayer (2013)) are unsuitable in this context because they pass censoring points through non-linear functions. I compute standard errors by bootstrapping country pair observations across both stages. I do not resample translation rate data, as I view the fixed effect regressions to net source language effects and country-specific wages from nominal translation rates as a data construction step.