No 2015-13 – July

Working Paper

CEPII

Foreign Language Learning: An Econometric Analysis

Victor Ginsburgh, Jacques Melitz & Farid Toubal

Highlights

- An econometric analysis of learning foreign languages in all parts of the world.
- New dataset that covers 193 countries and 13 important languages.
- Trade deserves special emphasis on the influence of foreign language learning.

RESEARCH AND EXPERTISE
ON THE WORLD ECONOMY

Abstract

The paper is devoted to an econometric analysis of learning foreign languages in all parts of the world. Our sample covers 193 countries and 13 important languages. Five factors significantly explain learning: the world population of native speakers of the home language, literacy, the world population of speakers of the target language, trade with foreign speakers of the target language, and the linguistic distance between the home language and the target language. All five factors affect the broad decision to learn but the last three also point to the choice of the particular language to learn. The world population of speakers of the native language discourages learning in general while literacy promotes it in general. Instead, the world population of speakers of a specific target language and trade with speakers of the specific language prompts learning of that language while the linguistic distance between the home and the foreign language discourages learning of that language. Trade may well deserve special emphasis, not only for its quantitative effect, but also because its direction can change faster and by a larger order of magnitude than the other factors. Controlling for individual acquired languages, including English, is of no particular importance.

Keywords

Language learning, Language and trade, English as a global language, Linguistic distance.

JEL

F10, F20, Z00, J00.

Working Paper

CEPII

CEPII (Centre d'Etudes Prospectives et d'Informations Internationales) is a French institute dedicated to producing independent, policy-oriented economic research helpful to understand the international economic environment and challenges in the areas of trade policy, competitiveness, macroeconomics, international finance and growth.

CEPII Working Paper
Contributing to research in international economics

© CEPII, PARIS, 2015

All rights reserved. Opinions expressed in this publication are those of the author(s) alone.

Editorial Director:
Sébastien Jean

Production:
Laure Boivin

No ISSN: 1293-2574

CEPII
113, rue de Grenelle
75007 Paris
+33 1 53 68 55 00

www.cepii.fr
Press contact: presse@cepii.fr

RESEARCH AND EXPERTISE
ON THE WORLD ECONOMY

1. Introduction

Language learning in a multilingual world receives considerable attention by linguists, historians, philosophers, and social scientists. Economists also possess a game-theoretical literature with empirical application to the learning of world languages (Selten and Pool 1991, Church and King 1993, Shy 2001). Yet econometric work on language learning on the international level has lagged behind. The econometric work on learning has been largely confined to the decision of immigrants and linguistic minorities to learn the primary language in their country of residence in order to increase their work possibilities and wages.[1] To our knowledge, the only econometric study thus far of the learning of *foreign* languages (in common use abroad but not at home) is a paper by Ginsburgh, Ortuño-Ortin and Weber (2007) concerning the learning of English, French, German and Spanish in the European Union. Here we propose to refine their research in one respect and to extend it in another. We refine it by explicitly acknowledging the presence of at least one language in each country that is needed for daily living and is a dominant choice for any immigrant. To simplify the analysis, we apply the same idea to linguistic minorities, possibly concentrated in certain regions, like Basque speakers in Spain or Gujarati speakers in India. It is only the choice of learning other languages that concerns us. We extend their analysis by taking a world view of the subject and dealing with the learning of 13 important languages in 193 countries. These languages are Chinese, English, Spanish, Arabic, Russian, French, Portuguese, German, Malay, Japanese, Turkish, Italian and Dutch, in descending order of number of speakers. Our data is cross-sectional and centers around 2005. Despite the considerable research to date on the influence of common languages on foreign trade[2] and the wide awareness of the role of foreign trade in stimulating learning of foreign languages,[3] this is also the first econometric work thus far to study the impact of trade on language learning. We consider trade not only as an inducement to learn but also a major reason for the heterogeneity of learning decisions. Though the world distribution of speakers of languages is the same everywhere, trade with different parts of the world differs greatly by country and offers different incentives to learn various languages. Moreover, within any particular country, citizens engaged in trade with different parts of the world may take different decisions about which language to learn.

We examine five separate influences on learning based on elementary theoretical analysis and the availability of data, and all five of them emerge as important with the expected sign. The first of these influences is the size of the world population of speakers of a foreign language; it encourages learning. The second is literacy, which encourages learning too. Third, a large world population of speakers of the home language discourages learning. People who possess a large language have less incentive to learn any other one. Fourth, linguistic distances also discourage learning. When the distance between languages increases, learning decreases. Last, trade with speakers of a foreign language is an incentive to learn the speakers' language.

*The authors would like to thank the participants in seminars at CREST, Paris, Ekaterinburg, Russia, University of Jaume 1, Castelon, Spain, and ECARES, Belgium, for valuable comments. They are especially indebted to Julien Martin and Olivier Loisel for their comments. Victor Ginsburgh wishes to acknowledge the support of the Ministry of Education and Science of the Russian Federation, grant No. 14.U04.31.0002, administered through the NES CSDSI. Farid Toubal wishes to acknowledge the support provided by the iCODE Institute (Idex Paris-Saclay).

[1] Research on the benefits of such learning by immigrants and minorities goes back far and is sizeable. See the collected essays in Chiswick and Miller (2007) and the contributions of many others (Bratsberg and Nasir 2002, Dustmann and Van Soest 2001, 2002, Fry and Lowell 2003, Grin 1999, and Vaillancourt 1996).

[2] See Frankel (1997), Anderson and van Wincoop (2004), Melitz (2008), and Egger and Toubal (2015), among others.

[3] For survey evidence that confirms the interest of exporting and multinational firms in acquiring foreign language skills, see The British Chambers of Commerce (2003-2004), Feely and Winslow (2005), and Hagen et al. (2006). See also Ginsburgh and Prieto (2011).

There are obvious differences between the operations of the five influences. Literacy encourages learning without regard to the target language. The size of the home language discourages learning generally too. On the other hand, the size of the population of the target language, and trade with this particular population, both promote learning of the particular language. Finally, the linguistic distance between home and target language affects the choice of foreign language to learn.

The trade motive for language learning emerges as the most important factor in our empirical findings. On the basis of our results, conditional on the presence of learners of a language in a country, a one percentage point increase in the trade share with speakers of the language will increase learners of the language (as a percentage of the total population) by around 1.4 percentage points. This is a large effect. It emerges after controlling for the reciprocal effect of learning on the trade share; that is, after instrumenting the trade share. Without conditioning on positive learning, a one percent increase in the trade share with speakers will also increase the probability of some positive learning after controlling for endogeneity. A doubling of the trade share causes a 13% probability of some positive learning where there is none. This next effect is smaller than the aforementioned one but as precisely estimated. These are also estimates without controls for national specificities. If we control for such specificities by introducing a separate fixed effect per country, the influence of trade on learning even goes up. It rises from 1.4 to 2.1 percentage points for learning when there is some and from a 13% to a 16% probability of learning when there is none.

The paper proceeds as follows. Section 2 provides some theoretical discussion of the influences on learning that we choose to investigate in the empirical study. Section 3 discusses the econometric model. Section 4 turns to the data and section 5 describes the estimation method. Sections 6 through 8 are devoted to results.

2. The theoretical background

In explaining language learning on a world basis, the game-theoretical literature provides a useful point of reference. This literature would suggest treating the total numbers of speakers of the different languages in the world and the costs of learning as outstanding factors. To explain, assume that there is a single home language per country and that everyone in each country possesses this language. In each country, therefore, all people can already communicate with the whole world population of speakers of their home language, including those who live abroad. The larger this total world population is, the better able they are to communicate and the lower are their marginal benefits of learning another language. Therefore, the total world population of speakers of the home language exerts a negative influence on learning at home. This theoretical result accords well with common experience. We encounter more monolingualism in large-language countries, like the US or the UK, than in small-language ones, like the Baltic ones.

In the case of each country, consider next the total world number of speakers of a foreign language. The larger this number is, the larger are the benefits of learning the language. On this ground, Chinese and English should attract more learning than other languages since they are the two largest in the world. Arabic and Spanish should rank high too.

However, as the game-theoretical literature stresses (see Gabszewicz et al. 2011 as well as Selten and Pool 1991 in this respect), there are costs of learning, if only the time and effort required to learn. One international measure of this cost is the distance between the home and the foreign language. The specific measure of this distance we shall use rests on expert judgments by ethnolinguists. Use of this measure should clearly help to reconcile the evidence with the fact, for example, that there is less learning of Chinese, Japanese and probably Arabic in countries using Indo-European languages than sheer numbers of speakers would lead us to expect. Another indicator of the cost of learning is the literacy rate. Higher literacy, implying higher education, should make learning easier and thereby promote learning of foreign languages.

If we limit ourselves to the preceding influences on learning, however − namely, world numbers of respective speakers of home and target language, linguistic distance, and literacy − we miss one important dimension, if not several. Consider the relative interest of learning English, French and German in a Spanish-speaking country in South America as opposed to Spain. For both countries, world speakers of these languages and the linguistic distances between them are exactly identical[4] and differences in the literacy rate cannot explain a preference for learning one language rather than another. Yet it is evident that French and German are of greater interest relative to English in Spain than South America because of higher levels of interaction with French- and German-speaking countries in Spain. Introducing bilateral trade ties between countries should evidently help to repair this difficulty. We know that bilateral trade ties reflect not only commercial ties but also geography and possibly history, for example, a common ex-colonial past. The impact of this variable is multi-faceted. From the standpoint of people as consumers, trade ties are an additional reflection of the benefits of learning foreign languages besides the world number of speakers. These ties perhaps focus more on commercial interests than the world number of speakers does (even allowing for trade in cultural products and tourism). Thus, in the presence of trade ties, the total world population of speakers of a target language is possibly best interpreted as mainly reflecting the non-market advantages of learning this language, stemming from the ability to interact socially with foreign speakers and to benefit from their cultures and cultural heritages. From the standpoint of people as producers, trade ties reflect the costs of learning. For a person who is professionally engaged in foreign commerce, trade with speakers of the target language reduces the opportunity cost of the time needed to learn the language because of more frequent, more effective and better motivated occasions to practice and because of a higher value of the time spent learning (via higher wages and profits).

A bit of notation is useful next. Let J be the target language and K be the home language in country K where language $J \neq K$. Next, let N_J be the world population of the target language, N_K be the world population of the home one, T_{JK} be the trade of country K with the N_J population of the world, D_{JK} be the linguistic distance between languages J and K and I_K be the literacy rate in country K. Define α_{JK} as the share of the population in country K that learns language J. Based on our theoretical discussion, the general function we propose is:

$$\alpha_{JK} = F\,(N_J, N_K, T_{JK}, D_{JK}, I_K), \text{ for all } J \text{ and } K,\ J,\ J \neq K \qquad (1)$$

with $F'(N_J) > 0, F'(N_K) < 0, F'(T_{JK}) > 0, F'(D_{JK}) < 0$ and $F'(I_K) > 0$.

Evidently T_{JK} needs more precision. We shall define T_{JK} as the ratio of the total trade of country K with the J-speaking world; namely:

$$T_{JK} = \frac{\sum_{c\in C} \sigma_{cJ}\, BT_{cK}}{\sum_{c\in C} BT_{cK}} \qquad (2)$$

where C is the set of country K's trading partners, σ_{cJ} is the percentage of speakers of language J in country c, BT_{cK} is the total trade of country K with country c, and the denominator in eq. (2) is therefore country K's total trade. This choice of specification has two advantages. First, T_{JK} is a fraction (as is α_{JK}), therefore a pure number. Next, T_{JK} reflects competition between languages. T_{JK} can only rise at the expense of trade with speakers of other languages than J. This makes sense since competition between languages is a fact of

[4] This is strictly true, of course, only on our present assumption of 100% Spanish speakers in both countries.

life. The time spent on learning one language cannot be spent on learning another and people's total time and capacity to learn foreign languages are limited.

Two further points need attention. First, our decision to disregard the learning of the home language puts special importance on the definition of this language. Many countries possess large minority languages. While we disregard the decision of a German resident to learn German in Germany, we do consider the decision of German residents to learn Turkish even though there are numerous native Turkish-speakers in the country. Likewise, we consider the learning of Spanish in the US though there are millions of native Spanish-speakers in the country. The need to draw a clear line on this issue demands some hard choices and leads us to refer principally hereafter to a "dominant" language rather than a "home" language. A "dominant" language gives less the impression of a language that is spoken by everyone or necessarily the overwhelming majority. Importantly in this respect, we shall also engage in some robustness tests about our choices of dominant language. Moreover, we shall recognize two dominant languages in some countries. Second, we ignore a central aspect of the game-theoretical literature on language learning: namely, that the learning of the dominant language in a country by foreigners diminishes the welfare benefit of learning the foreigners' language by speakers of the dominant language since it now becomes possible to communicate with the learners at no extra cost. However, foreign learning decisions may not weigh heavily in decisions to learn foreign languages, especially if the decisions are decentralized. In addition, some of the gains of learning depend on the ability to understand what others say and write in their own language. We shall assume that the effect of foreign decisions on current ones at home is negligible.[5]

3. Econometric specification

We shall test a linear world approximation to eq. (1): namely,

$$\alpha_{JK} = \beta_0 + \beta_1 N_J + \beta_2 N_K + \beta_3 T_{JK} + \beta_4 D_{JK} + \beta_5 I_K + \varepsilon_{JK} \tag{3}$$

The test requires that the five right-hand side variables be exogenous or independent of α_{JK}. This exogeneity can reasonably be accepted for linguistic distances and literacy rates; but it cannot be for the rest. In the case of N_J and N_K, the problem is easily repaired. Both variables will be highly correlated with the corresponding ones for native speakers and we shall therefore measure them on the basis of world native-language populations rather than world speakers. The difficulty is far greater for T_{JK}. We can rest T_{JK} exclusively on native speakers too by measuring σ_{cJ} accordingly in eq. (2) and we will do so. This helps but it cannot suffice since knowing a language promotes trade with native speakers as well as other speakers. We must therefore go further and instrument T_{JK}. Our choice of instrument is taken from Frankel and Romer (1999), who faced a similar problem to ours. They needed ratios of trade to GDP that were independent of economic growth; we need T_{JK} values that are independent of language learning. Their solution, now widely accepted with some reservations that do not concern us (as we will shortly explain), was to base trade values strictly on "geographical" characteristics (in their terms) such as national land area, status as landlocked, and population size. We will repeat except for adding an extra step.

In the first step, we estimate a bilateral trade equation between countries *c* and *K*, as they do, where

$$ln\ (BT_{cK}/GDP_K) = \alpha_o + \alpha_1\ ln\ D_{cK} + \alpha_2\ ln\ P_c + \alpha_3\ log\ A_c + \alpha_4\ log\ P_K + \alpha_5\ log\ A_K + \alpha_6\ (L_c + L_K) + \alpha_7$$

[5] This is in no way to deny that political conflicts and issues of coordination about languages are of first order importance within countries and within international political organizations. See Laitin (1994) and Pool (1991, 1996) and also Ginsburgh and Weber (2011).

$B_{cK} + \varepsilon_{cK}$ (4)

As in eq. (2), BT_{cK} is the bilateral trade of country K with country c, GDP_K is the gross domestic product of country K, P is population, A is land area, L is a dummy for landlocked countries and B is a dummy for common border between K and c.[6] Next, we obtain the exponential of the estimates (or estimates of BT_{cK}/GDP_K), as they do, which we label $\hat{B}_{cK}$. Following, we sum the ratios $\hat{B}_{cK}$ over the entire set C of K's trade partners. This is as far as Frankel and Romer go. But we go on to calculate a weighted sum of these last ratios with weights depending on the partners' respective ratios of native speakers of language J, namely, σ_{cJ}, and finally we construct the ratio of the weighted to the un-weighted sum:

$$\hat{T}_{JK} = \frac{\sum_{c \in C} \sigma_{cJ} \hat{B}_{cK}}{\sum_{c \in C} \hat{B}_{cK}} \qquad (5)$$

$\hat{T}_{JK}$ will be our instrument for T_{JK}. It is evidently an estimate of the share of country K's trade with native speakers of language J. The basic difference from FR is the addition of native-language weights. However, since learning of language J in country K will not affect *native* speakers of language J in country c, α_{JK} cannot affect our instrument any more than economic growth can affect theirs. The only criticism that Frankel and Romer ever faced is that the impact of their instrument on growth did not necessarily reflect the strict influence of trade (see Rodriguez and Rodrik 2001, and Noguer and Siscard 2005). But this criticism does not concern us since it would not particularly worry us if the effect of $\hat{T}_{JK}$ on α_{JK} reflected other correlated interactions between countries besides trade since in our work T_{JK} is simply a general control for international interactions between countries in the presence of the other influences on α_{JK}.

We shall therefore estimate eq. (3) after instrumenting T_{JK} by $\hat{T}_{JK}$ and after substituting native-language series for N_J, N_K and σ_{cJ}. Because N_J and N_K are worldwide values and may go from over a billion (for Chinese) to very small values for a language like Wolof (important in Senegal) or Inuktitut (Greenland), we shall express them in logs.[7] The other variables can be left as they stand. Indeed, α_{JK}, T_{JK} and I_K are national shares while distances D_{JK} are normalized on the unit segment and every impact on α_{JK} will be easy to interpret.

Eq. (3) recognizes no lagged effects even though learning languages takes time. In this regard, note that we only possess data for language learning for a single date. The data pertains to the net cumulative learning over the entire past approximately in 2005 and we cannot deal properly with the dynamics. We could, of course, have admitted a lag in the influence of T_{JK}, our one explanatory variable that possibly moves rapidly over time (though less so than it might appear, since T_{JK} depends on the linguistic structure of national trade rather than the level of national trade while this structure is likely to be more stable than the level). Yet how slow is language learning in practice? People forget languages through

[6] In fact, Frankel and Romer add interaction terms for common border with each of the other six variables on the right. We have used them too, though we do not display them above, as Rodriguez and Rodrik (2001) and Noguer and Siscart (2005) have in discussing FR's work (though Irwin and Tervio 2002 show that these interaction terms do not matter in applying the same test as FR's to earlier twentieth-century cross-sections).

[7] It would make no difference if we took logs of the ratios of N_J and N_K to world population: the estimates would be the same.

disuse and may never have been able to converse in foreign languages they studied in school. Yet a year or two may suffice to learn a language with adequate motivation and occasion to practice. In light of the arbitrariness of any imposed lag structure, we will provide estimates without lags even though we investigated the impact of lags.[8]

Two control variables also readily come to mind. One is a dummy variable for ex-political administration or ex-colonization of country *K* by a foreign country with native language *J* since 1939. A former member of the Soviet Union is more likely to speak Russian, and a former British colony is more likely to speak English. The second control is a dummy variable for Indo-European languages. Among the 13 destination languages in our study, eight are Indo-European, while the other five – Chinese, Arabic, Malay, Japanese and Turkish – all belong to different language families. This may matter for several reasons. Indo-European languages are geographically concentrated in Europe and the Americas and familiarity may therefore make it easier to learn one for those who possess another (the more so if both belong to the same branch of the ethnological language tree, like English, German, Dutch, or French, Italian, Spanish, Portuguese). Learning a third language may also be easier for those who already know a second. Finally, except for Russian, the eight Indo-European languages use the same alphabet. The introduction of linguistic distances may not adequately reflect these factors. It could thus be that an Indo-European dummy would have a positive effect.

4. Data

The necessary data require a table with columns representing our 13 destination languages and rows for our 193 countries. Each cell of the table contains the number of individuals (or their share in source country *K*) who speak each of the 13 destination languages *J*. Searching for these numbers can proceed in three ways. In some cases (the European Union or EU in particular), we were able to work by row (which of the 13 languages are spoken in, say, Spain). In many other cases, we had to proceed by column (in which countries do people speak Spanish). Most often, we had to combine both approaches, making sure that our figures are consistent.

For most spoken and native languages in Western Europe, we proceeded by row (source countries), using the EU survey *Special Eurobarometer 243* (2006), which covers the current 28 EU members plus Turkey and includes 32 languages, 25 of which are part of N_K . In recording the data we added answers to the two following questions: “What is your maternal language” and “Which languages do you speak well enough in order to be able to have a conversation, excluding your mother tongue (… multiple answers possible).”

For countries other than members of the EU, we completed the table using a wide variety of sources, mostly proceeding by column (destination language):

– For English, we relied mainly on Crystal (2003a, p. 109) for the rest of the world outside of the countries in the EU survey. Because of the rapid ascension of English as a world language in our study period, we suspect the main flaws in our series to be some of the zeros for spoken English (for example, in South Korea).

[8] We experimented with lagged average values of T_{JK} for the decades of 1990, 1980, 1970, 1960, and 1950 while instrumenting with these values for T_{JK} and got similar results. Once we get back before 1990, Russian virtually disappears from the trade data because of the ex-Soviet Union, and the number of countries also drops progressively. By 1950, the number of countries is down to 127 (from 193 in 2005) in the full sample and to 59 (from 94 in 2005) in the positive sample. It might also seem that lagging T_{JK} or else using the lagged values of T_{JK} as instruments would have helped to handle the problem of simultaneity. But this would be a mistake. The lagged values still depend on language learning in the past, for which we have no separate data and that would be highly correlated with T_{JK} in 2005.

– For French, we used mainly the “Estimation du nombre de francophones dans le monde” website http://www.axl.cefan.ulaval.ca/francophonie/OIF-francophones-est2005.htm.

– For German, we relied mainly on *Ethnologue*.

– For Spanish, we used an unusually well documented Wikipedia website, with many dozens of references to official sources, http://en.wikipedia.org/wiki/Spanish _language.

– For Portuguese, we used a website entry for “Geographical distribution of Portuguese” that was no longer available on the web when we last checked in December 2013.

For other languages, we relied heavily on web searches, first, by language (columns), next by country (rows) in *Ethnologue*. While this source of information is extensive for native languages (L1 in *Ethnologue*), it is far less so for spoken language by non-natives (L2), where data appear on a selective basis (though the source remains important). Therefore, we made further web searches for L2 for the 13 languages in our study. In particular, in the case of Russian, we exploited a Gallup poll of non-EU members of the ex-USSR from a website titled “Russian language enjoying a boost in Post-Soviet states” (http://www.gallup.com/poll/ 109228/russian-language-enjoying-boost-postsoviet-states.aspx). Arabic was a particular problem. For lack of a better solution, we made numerous inferences about L2 from literacy rates in Arab-speaking countries. In identifying languages, we assumed Tajik and Persian (Farsi) to be the same language, and did the same for Hindi and Hindustani, Afrikaans and Dutch, Macedonian and Bulgarian, Belarusian and Russian, Icelandic and Danish, Turkmen, Azerbaijani and Turkish, as well as Zulu and Xhosa.

In general, our two outstanding sources are the EU survey *Special Eurobarometer 243* and *Ethnologue*.

The dependent variable in our model, α_{JK}, is the ratio of non-native speakers of language J in country K to the number of inhabitants of country K. The 13 N_J values follow directly from the world values of native speakers in levels while the N_K values vary by country depending on its native languages.[9]

Table 1 provides information about the 13 destination languages. It lists the total number of people who use them as mother tongue in column 2, the number of worldwide speakers in column 3. Column 4 contains the ratio of worldwide speakers to native speakers (“the language multiplier”). Malay, an official language in Malaysia, Singapore and Brunei, has spread throughout Indonesia where it became a *lingua franca*, and has the largest multiplier. French comes second and is moderately ahead of English. The language is widely spoken in many former French colonies and overseas territories particularly in Africa where native speakers are few. German and Dutch (which is spoken in The Netherlands, Belgium, parts of the Caribbean and a variation of which, Afrikaans, is an official language in South Africa) come next. Japanese, Chinese and Portuguese (mainly spoken in Portugal and Brazil but little elsewhere) close the list.

Our choice of a primary language for each country is important. It affects both the learning decisions we drop out and the definitions of the distances D_{JK}. In most cases, this language is obvious and can be identified with the native language of the majority, such as German in

[9] To be precise, N_K is the sum of the world values of the country’s native languages multiplied by the respective percentages of the native speakers of these languages within the country. Take a simple example of a country with 60% native speakers of language A and 40% native speakers of language B. For this country, N_K will be equal worldwide native speakers of language A times 0.6 plus worldwide native speakers of language B times 0.4. In total, 106 different languages enter in the determination of N_K over the 193 countries. Footnote 13 contains further detail.

Germany. Yet this is not always as easy. For example, in India, Hindi and English are both widely spoken, and we decided to treat both as primary home languages. In all, there are 21 cases of this sort (which will be mentioned below). In another set of ten cases, always associated with high linguistic diversity, the problem is not so much to choose between two languages but to pick a single one. Invariably, however, one major world language receives official status and we consider this language to be the one whose learning falls outside of our analysis. Seven of these instances concern French (Burkina Faso, Democratic Republic of Congo, Central African Republic, Guinea, Republic of the Congo, Senegal and Togo), two concern English (Northern Mariana Islands and Sierra Leone) and one Portuguese (Guinea Bissau). We could have assumed that no home language exists at all in these cases, but we chose to stick to the principle that in every country there is at least one particular language, if not two, the acquisition of which dominates the rest for permanent residents who do not already possess it (or one of the two).

A number of different cases can be distinguished.

(a) Countries with a primary language that does not belong to the 13 destination languages provide 13 observations, since their inhabitants can decide to learn any of the 13 languages, though many α_{JK} will equal zero. The same will be true in four of the 21 cases of countries with two primary languages because neither of them belongs to the destination languages. This is so for Afghanistan (Pashto and Persian), Bhutan (Djonkha and Nepali), Bosnia and Herzegovina (Bosnian and Serbo-Croatian), and Fiji (Hindi and Fijian).

(b) Countries (such as Germany, Saudi Arabia or Russia) whose primary language is one of the destination languages provide 12 observations at most, since their acquisition by residents of these countries is not taken into account.

(c) In nine of the 21 cases with two primary languages such as India, only one of them is relevant and there are still 12 observations. This is so for the Cook Islands (Maori and English), India (Hindi and English), Nauru (Nauruan and English), Niger (Hausa and French), Nigeria (Hausa and English), Niue (Tonga and English), Palau (Palauan and English), the Philippines (Tagalog and English) and South Africa (Zulu and Dutch).

(d) In eight cases with two primary languages, both belong to the 13 destination languages, and there are only 11 observations. These eight cases are: Aruba (Spanish and Dutch), Cameroon (French and English), Chad (Arabic and French), Djibouti (Arabic and French), Mauritius (French and English), Singapore (Chinese and English), Suriname (Dutch and English), and Vanuatu (French and English). Note that we do not regard Belgium, Switzerland or Canada as belonging to these cases despite the regional significance of French as a second national language in all three. However, we will engage in a robustness test on this issue.

The primary language also serves to define the distance D_{JK} between the source and the destination language. The distances come from the Automated Similarity Judgment Program or ASJP, an international project headed by ethnolinguists and ethnostatisticians (see Brown et al 2008). As of late 2010, when we got access, the ASJP had a database covering the lexical aspects (word meanings) of close to 5,000 of the world's nearly 7,000 languages (Bakker et al. 2009).[10] The ASJP values go from 0 (no distance) to 105 and were normalized on the unit segment. In the case of two primary languages in a country, we weigh the two distances, mostly but not always half and half.[11]

[10] See also http://wwwstaff.eva.mpg.de/~wichmann/ASJPHomePage.htm

[11] For example, for India, we weigh Hindi .67 and English .33.

The advantage of this source is that linguistic distances are not restricted to Indo-European languages (as they are in Dyen et al 1992) and yet were computed by ethnolinguists (based on a tradition that goes back to Swadesh 1952). There is an alternative measure of linguistic distance suggested by Laitin (2000) and Fearon (2003) that has become popular recently and that founds the distances on the *Ethnologue* classification of language trees. However, we prefer our measure in two respects. The Fearon-Laiton measure always supposes maximal distance between languages belonging to different trees. Further, the measure assumes that a distance of 0.5, for example, means the same in the Indo-European group as in the Altaic one. The ASJP measure avoids either assumption.[12]

Trade shares T_{JK} required converting a K by K matrix of bilateral trade values into a K by J matrix of country shares of total trade with all native speakers of language J in the rest of the world. To proceed, we multiplied K's bilateral trade with each of its trade partners by the respective percentage of native speakers of language J in the partner country, summed over all partner countries and divided by the total trade of country K (see eq. (2)). Bilateral trade series come from the BACI database of CEPII (which corrects for various inconsistencies; see Gaulier and Zignano, 2010). GDP and population data come essentially from the Penn World Tables, literacy rates from the CIA World Factbook and ex-colonial relations from Head, Mayer and Ries (2010). The series for the right hand side variables in eq. (4), such as land area, common border or landlocked countries are taken from Mayer and Zignago (2011). The base year for most data is 2005, though language data cannot be constructed for any single year on a world basis and refers to different years between 2001 and 2008. The same problem exists for literacy rates, a slow-moving variable, which we based on recent data.[13]

Table A1 of the Appendix provides summary statistics for our main variables.

5. Estimation method

The total number of observations is 2,365 (less than 193 times 13 or 2509 for reasons that follow from the preceding section), though there are only 240 with non-zero left-hand side values α_{JK}. There are two basic reasons for the predominance of zeros. Each individual learns a small number of languages at best. This can account for many zeros, even at the national level. Second, and probably more important, we only collect values of α_{JK} that are at least equal to one percent at the national level. It does not appear reasonable to suppose that a single mechanism determines the numerous zeros and the wide array of positive values when learning takes place. Therefore we provide two separate estimates of the basic model. First, we consider the binary choice between learning and not learning for the full sample and estimate the model using probit. In this case, each parameter estimate can be read as the rise or fall in the probability of learning that results from a change in the associated variable of one percent. Next, we consider the percentage of learners conditional on positive learning (240 observations) and apply ordinary least squares. In this case, the appropriate interpretation of each individual parameter needs no further clarification. In both cases we instrument for trade, therefore using probit with instrumentation in the former and two stage least squares in the latter.[14] However, to allay any lingering doubts about the

[12] Notwithstanding, we experimented with the Fearon-Laitin measure of D_{JK} as well as the ASJP one (as Melitz and Toubal 2014 had in a study of bilateral trade). The results are similar (as they were in their case).

[13] We were unable to retrieve population and/or output data for 2005 in a small number of cases (Anguilla, British Virgin Islands, the Falklands), and replaced them with data for years close to 2005 based on web searches.

[14] In similar situations, researchers sometimes propose a third estimate concerning the probability of positive learning based on the combination of the two estimates (see Wooldridge 2002, pp. 536-538, Wooldridge 2007, p. 573, and for a relevant Stata command and associated discussion, Belotti et al. 2012). However, in all of the examples (which sometimes refer to "two-part models"), there is no endogeneity in the explanatory variables and therefore no need for instrumentation. The missing third estimate does not strike us as a fundamental absence.

zeros, we also furnish results of estimates of the basic model based on the treatment of the data as a single sample in Table A3 of the appendix.

6. Main estimation results

Our main results are presented in Table 2. The probit estimates in the first three columns, all based on the full sample, are the marginal effects evaluated at the sample means of the variables. As the first column shows, all five explanatory variables are highly significant with the expected signs prior to any correction for the endogeneity of trade. The second column gives the first stage of the IV probit and shows that the instrument for trade is strong. In the third column, we see that once we correct for the endogeneity of trade, all five coefficients notably drop but remain significant. Based on the estimates, the largest effect by far on learning appears to be trade. Specifically, there is a 13% probability that a doubling of trade will result in some learning of the destination language. If we look at the standardized "beta coefficients" instead (Goldberger 1964, pp. 197-200), the coefficient of trade (0.25) is not higher at all than three of the other four significant ones: the negative ones for world speakers of the native language (-0.28) and linguistic distance (-0.23) and the positive one for literacy (0.36). Yet trade is also more variable than these other three influences, especially linguistic distance, a constant, and the literacy rate, often close to one. Thus, the emphasis on trade remains perhaps right.

Columns (4) to (6), concerning the positive sample, deal with the results conditional on positive learning. The population of native speakers of the home language, trade and linguistic distance remain highly significant (linguistic distance below the 99% level) in the estimate without correction for endogeneity (column 4), though the world population of speakers of the target language ceases to be significant at conventional levels and literacy becomes totally insignificant. Once again the instrument for trade performs well (column 5). After correction for endogeneity (column 6), the significance of all the variables remains the same except that the world population of speakers of the target language now becomes totally insignificant. In addition, the coefficient for trade is substantially higher than before in column 4. A one percentage-point increase in the ratio of trade with native speakers of the destination language would increase learning of the language by 1.4 percentage points, conditional on positive learning. This effect is much stronger than the two other significant ones (and in this case it does not much matter if we look at the standardized "beta" values of the coefficients instead). The negative significant effect of native language on learning is also of some consequence. A 100 percent increase of speakers of this language would reduce learners of other languages by 2.5 percent. Thus, in a nation of 50 million native speakers in which there are already learners, this would mean a reduction of 1.25 million learners.

In Table A2 of the Appendix, we also show the results of alternative estimates of our model based on the treatment of the data as a single sample. In this case, we use IV Poisson, IV negative binomial and IV fractional logit. The results are remarkably similar with all three quasi-maximum likelihood estimation methods and they also yield the same signs and significance of the variables as those in our basic two-part model. However, as we underlined earlier, we reject the idea of a single parameter estimate for learning independently of the presence or absence of learners in a sample where 90 % of the values are zero. Therefore, we stick to our two-part model for the rest of the work.

There remains the curious fact that the correction for endogeneity reduces the estimate of the influence of trade in the full sample but increases it in the positive sample. The reduction in the full sample is the easier one of these two results to interpret. Suppose, as expected, that learning increases trade with speakers of the destination language rather than the opposite. If so, the one-stage probit estimate is biased upward and, by removing the bias, the correction for endogeneity in the instrumental-variable probit estimate should yield a lower estimate for trade. In the case of the positive sample, the same reasoning holds but there are two basic differences. First, the number of observations is much smaller and second, the distinctions are finer. On both counts, observation errors in the measure of trade may be more important than in the full-sample estimate. These errors tend to bias the estimate of the influence of trade downward rather than upward. Assume that this negative bias trumps the positive one from the reciprocal effect of learning on trade so that the OLS estimate is biased downward. The rest of the reasoning is more special. Suppose, further, that observation errors in the trade variable are particularly important in estimating the effect of learning when learning has a reciprocal positive effect on trade. In other words, the errors in the measure of the trade variable are far more correlated with learning before than after correction for reciprocal effects. In this case, the corrected estimates of the influence of trade would be closer to the truth and higher. This is the fundamental explanation we see for the higher 2SLS than OLS estimate in the positive sample.[15]

7. Robustness tests

We performed seven basic robustness tests.

The first introduces ex-colonial languages and Indo-European languages as controls. Since the results of adding each control separately hardly matters, we simply show the results of adding both jointly. As seen in Table 3, adding both controls changes little. The basic variables are only modestly affected. Both of the new variables are significant in the full sample and not in the positive sample after correction for endogeneity. It would thus be possible to retain the two; but the baseline model is satisfactory.

The next two robustness checks cope with a couple of data issues. Two of our 13 languages, Chinese and Arabic, are "macrolanguages" in *Ethnologue*’s terms; they bundle native speakers of distinct and often mutually unintelligible dialects. The two represent single languages only by virtue of custom and the tendency of native speakers to identify themselves with the general label. Mandarin serves as the main reference point for Chinese, Standard Arabic for Arabic. Because this can lead to doubts, we performed tests ignoring one or the other or both. Table 4 shows that there is hardly any noticeable change.

The next issue concerns the possibility that our data for spoken English are too low since, as Table 1 shows, they yield a total of around 1.1 billion speakers worldwide, whereas a higher figure of 1.5 billion based on a global approximation by Crystal (2003b, pp. 68-69) circulates widely. This last estimate has been repeated on the prominent website of the British Council. In fact, we predominantly repeat the same figures for individual countries that Crystal provides, which cover only 75 “territories where English has held and continues to hold a special place” (2003b, p. 60), by which by and large he evidently means territories that were under the administrative control of English-speaking powers at some time in living memory or else where the language is official or both. Those figures therefore do not include spoken English in places like the Netherlands, Germany and the Scandinavian countries where it is widely spoken but has never been either the language of the ruling political power or official. Upon close examination, Crystal’s large global number of speakers (which he offers in a very circumspect manner) must come from much higher figures than ours in parts of Asia. Kachru (2010, p. 207), whose earlier work Crystal cites, produces a table for “Asia’s English-using populations” which contains roughly 200 million more Chinese English speakers than our

[15] Compare this analysis with that of Frankel and Romer (1999) who faced the similar problem of explaining why the estimates of the impact of trade on growth were higher once they corrected for the reciprocal effect of growth on trade. One alternative possibility they entertain is an accident of sampling. But in a study of earlier historical samples, Irwin and Tervio (2002) show that this possibility is unlikely.

figure of 11 million and 100 million more (non-native) Indian English speakers than our 200 million (for India see also Crystal 2003b, pp. 46-49). Adding these numbers to ours would bring our total for English speakers to 1.4 billion. The rest of Kachru's numbers resemble ours and are sometimes even lower. We added these two figures for India and China in our data. The change for India cannot make any difference, since we regard English-learning in India as domestic learning (and the 100 million added Indian speakers also do not alter N_J and N_K for the country, as those numbers rest on native speakers). We therefore experimented simply with an added 200 million English speakers in China. There is almost no change in the estimates, which we do not report here.

The next three robustness checks are concerned with more conceptual issues.

First, as emphasized, our trade variable focuses on relative trade in different languages. Yet trade could also have an effect on the incentive to learn languages across-the-board. It is not obvious that our trade variable would fail to reflect the common influence in this case. But notwithstanding, we experimented with adding the ratio of trade to output (a measure of openness) as a separate factor. We did so by introducing the variable as such or else also adding the product of the variable and world population of the destination language N_J (always using logs for the product but not necessarily for openness). Table 5 shows the outcome with openness alone (in logs). The coefficient is not significantly different from zero and its presence has virtually no effect on the other coefficients. The result is the same regardless which variant we use. We therefore conclude that T_{JK} by itself adequately reflects the influence of trade on language learning.

Secondly, in our previous estimates, we chose to treat the learning of the native language of some large minorities (for example, French in Belgium and Russian in Latvia) as the learning of a foreign language. These are debatable cases. Suppose instead that we define languages as "primary" if the native-language population represents 20 percent or more of the total population in a country. 14 extra observations now drop out of the analysis (including those for French in Belgium and Russian in Latvia) since the relevant languages become primary ones and domestic conditions therefore set the decision to learn them.[16] As can be verified in Table 6, the loss of these observations has almost no effect except on the trade variable, whose coefficient in both samples drops by nearly a third but remains highly significant.

The third robustness check responds to a diametrically opposite concern to the one motivating the previous check: the possibility that we may be wrong to ignore the domestic learning of the primary language at home by immigrants and minorities, and that the same principles should apply to their learning decisions as well. Including domestic learning (that is, the learning of German by Turkish immigrants in Germany, etc.) increases the number of observations by 137, of which 105 are positive.[17] This represents almost a 50 percent increase in the number of positive observations (345 instead of 240). There are also 32 extra zeros (besides the additional 105 positive values) for learning in the full sample. These are instances of no learning of our 13 languages despite the fact that they are primary. Results are shown in Table 7. The quality of the estimate drops drastically in the positive sample estimate. In addition, trade and literacy both perform more poorly. Trade remains significant

[16] The 14 observations are Russian in Kazhakstan (41 percent native), Spanish in Belize (36), French in Belgium (35), Spanish in Andorra (35), Russian in Ukraine (29), Italian in Malta (28), Russian in Kyrgystan (27), Russian in Latvia (26), Spanish in Gibraltar (26), French in Canada (23), Arabic in Israel (21), French in Switzerland (20), Turkish in Iran (20) and Turkish in Cyprus (20). In the positive-sample estimates, we lose only 12 observations since there is no learning of Arabic in Israel (despite the 21 percent of native speakers) or Turkish in Cyprus (despite the 20 percent of native speakers).

[17] Why not 144 more observations, which would bring the total up to exactly 13 times 193 or 2509? The reason is that there are seven cases where learning is impossible because we recorded 100% for native language: British Virgin Islands (English), El Salvador (Spanish), Montserrat (English), Portugal (Portuguese), Russia (Russian), Saint Pierre et Miquelon (French) and Turks and Caicos Islands (English).

in the full sample only below the conventional 95 % confidence level and becomes totally insignificant in the positive sample. Literacy also loses significance in the full sample and even acquires an implausible negative sign close to the 90 % confidence level in the positive sample. These adverse results are easy to interpret from our perspective. If domestic trade has a high priority over foreign trade in the learning decision of residents who lack the domestic language, it is not surprising that the performance of the trade variable in Table 7 would drop. Further, if learning the language in everyday use is the dominant choice for those who lack it, literacy might well be expected to be less important. We conclude that the additional observations in the table do not properly belong in the analysis and that the decision to learn the primary language of a country by immigrants and other permanent residents is indeed a subject requiring separate analysis: the incentives to learn are different.

Our last robustness test, Table 8, is a particularly strong one. We add country fixed effects. Consequently, speakers of native languages and literacy drop out since both variables are country-specific. An additional 191 country fixed effects enter. There is a remarkable stability in the results for the remaining three explanatory variables. In fact, these results are superior in one important respect: the positive effect of the world population of target languages emerges as highly significant with a positive sign in the positive-value sample. In addition, the trade effects are now higher while their significance is little affected. If we compare the new results with our baseline in Table 2, a one percentage point increase in the trade share with speakers of a foreign language increases the learning of the language by 2.1 instead of 1.4 percentage points (positive sample), and a doubling of the trade share with speakers increases the probability of learning by 16 instead of 13% (full sample).

8. Individual languages, or are some destination languages different?

Thus far we have also assumed that the same model holds for all 13 destination languages and that no special attention to individual languages is required. Accordingly, we have applied a common coefficient to the world population of native speakers of the destination language, regardless of the language, via N_J. A possible alternative is to introduce a separate interaction term for each language by multiplying a dummy for the language by N_J, the number of native speakers of the language, or simply, a dummy for each language (thereby ignoring the fact that some destination languages are larger than others). In both cases, the individual coefficients turn out insignificant, either separately or jointly.

In light of this negative result, what we show instead in Table 9 are the means and standard errors (as well as the *t*-statistics) of the residuals of the regressions in columns (3) and (6) of Table 2 for each destination language. This gives an idea of the direction of the residuals and how statistically significant they are. There is nothing to show for Japanese for the positive-value sample since there is no learning of that language in our database. There is also no standard deviation of the residuals in the full sample for Portuguese for which we have only one positive value (learning of Portuguese in Spain).[18]

As Table 9 shows, 11 of the means in the full sample are negative and in 10 cases (omitting Japanese) they fail to capture some positive learning, but none of them is even remotely significantly different from 0. In the positive sample, only the Chinese mean is highly significantly different from 0, but this result applies strictly to Malaysia and Singapore, the only two countries with positive observations for learning of Chinese in our database. The standard deviation is therefore based on only two residuals. Note also that the mean of the residuals for Chinese in the full sample, which takes into account all observations, is almost identical to the one in the restricted sample. Yet the former is totally insignificant because of

[18] The other positive values for Portuguese in our sample are for countries where the language is a primary one and therefore fall out of our analysis.

a much larger standard deviation.

The general impression from Table 9 is that the model performs in a similar way for all languages. One could say that English is the language that performs worst (mean error of –0.57 in the full sample). In addition, the mean error is negative (we under-predict), which can be interpreted to reflect the possibility (outside the confines of the model) that English is a world *lingua franca*, since there is more learning of the language than the model predicts in-sample. However, the standard deviation for this language is also by far the largest and denotes a significant percentage of cases of positive learning when the model predicts none (accordingly the *t*-statistic is low, 0.39). Furthermore, in the restricted or positive-value sample, the mean error for English is over-predicted and not particularly distinguishable from the rest (six of which are under-predicted, not counting Chinese). This goes against the idea of special status as a *lingua franca*. The result might seem contrary to a lot of independent evidence since English does indeed serve as a *lingua franca* in some limited areas like air traffic control, scientific writing and international sports. But the impression is contestable. When it comes to trade, the internet and publishing, other work shows that English does not require special treatment.[19] Our results therefore are not unusual. The case of Japanese deserves special mention too since there is no observation with positive learning for this language. Yet its mean residual of 0.2 with a *t*-statistic of 0.4 fits in well with the figures in the rest of the sample (as can be explained by linguistic distance).

9. Closing discussion

There is considerable interest today in the future linguistic map of the world, and particularly about how far English will go. The British Council has funded two important studies that were carried out by Graddol (1997, 2006) and speculation is wide. Crystal (2003b), Kachru (2010), Ostler (2010) and Huntington (1996, ch. 3) are noteworthy contributors to the issue. However, with the exception of Ostler, no effort was made to apply the same intellectual framework to other languages than English and in particular, no effort was made to use econometrics. Here we try to do both.

In our econometric modeling, we take as a point of departure Selten and Pool (1991) and Church and King (1993), except for dropping the interactions between learners. We also modify their analysis by adding trade as a factor, and we distinguish sharply between learning of foreign languages and the dominant language (or two languages) at home.

Our results, based on world data, support the view that a unified approach to language learning without any attention to particular languages has merit. International trade has a marked influence. The worldwide size of the native home language also influences learning of foreign languages, though in a negative way: if one’s home language is widely spoken in the world, there is less need to learn a foreign language. Linguistic distances have a negative effect on learning. Two other positive influences on learning show up, total world population of speakers of the destination language and literacy, though in both cases the influences are only clear for the decision to learn when there is none, not for additional learning when there is some. Finally, controlling for different languages does not help: once account is taken of our control variables, “all languages are equal.” If English is a separate factor as such, we could not find it. In the context of our research, this can be seen as a positive result, since it implies that learning English is subject to the same principles as learning other languages. It

[19] For trade, see Melitz and Toubal (2014); for the internet and publishing, see Melitz (2015). As regards translation (a branch of publishing), see Ginsburgh et al. (2013). On a different note, it might also seem, especially in light of the results for the full sample, that if we introduce a dummy for English alone, it would emerge as significant. But there is nothing special about English in this regard. Most of the languages emerge as significant in one test or the other (full sample or positive sample) when we introduce the languages alone, just as English does. We consider all such tests dubious and the right ones to be the sort to which we refer in the text and that we attempted, which admit as many different languages as possible simultaneously.

may therefore be wrong to try to assess the future of English in isolation, without allowing for similar incentives to learn other major world languages.

What can be said about the future of English? On the basis of our analysis, the evolution of trade will have a profound effect but its influence is complex. The effects of trade should be symmetric. Growth in Chinese/English trade should promote the learning of Chinese in native-English countries just as it should promote the learning of English in native-Chinese countries. Whether it will raise the importance of English relative to Chinese in the world will therefore depend heavily on the evolution of the share of trade with English speakers on the Chinese side relative to the evolution of the share of trade with Chinese speakers on the English side. That is what the econometric model shows.[20] The influence of demographic changes is simpler to analyze. Suppose for example that the Arabic and Spanish-speaking populations grow fast while numbers in the rest of the world remain constant. Then the Arabic and Spanish-speaking populations will wish to learn fewer foreign languages while speakers of other languages will wish to learn more Arabic and Spanish. Thus, Arabic and Spanish will become relatively more important, as Graddol (2006) foresees. Clearly, the basic demographic assumptions do not favor English.

References

Anderson, James and Eric van Wincoop (2004). “Trade costs.” *Journal of Economic Literature* XLII: 691-751.

Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichman, Cecil Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant and Erik Holman (2009). “Adding typology to lexicostatistics: A combined approach to language classification.” *Linguistic Typology* 13: 167-179.

Belotti, Federico, Partha Deb and Edward Norton (2012). “tmp: Estimating two-part models.” *The Stata Journal*, vv, ii: 1-13.

Bratsberg, Bernt, James Ragan and Zafir Nasir (2002). “The effect of naturalization on wage growth: A panel study of young male immigrants.” *Journal of Labor Economics* 20: 568-597.

Brown, Cecil, Erik Holman, Søren Wichmann and Viveka Velupillai (2008). “Automatic classification of the world’s languages: A description of the method and preliminary results.” *Language Typology and Universals* 61(4): 285-308.

Chiswick Barry and Paul Miller (2007). *The Economics of Language, International Analyses*. London and New York: Routledge.

Church, Jeffrey and Ian King (1993). “Bilingualism and network externalities.” *Canadian Journal of Economics* 26: 337-345.

CIA World Factbook. Available online at https://www.cia.gov/library/publications/the-world-factbook/.

Crystal, David (2003a). *The Cambridge History of the English Language*. Cambridge, UK: Cambridge University Press, 2d edition.

Crystal, David (2003b). *English as a Global Language*. Cambridge: Cambridge University Press, 2d edition.

Dustmann, Christian and Arthur Van Soest (2001). “Language fluency and earnings: Estimators with misclassified language indicators.” *Review of Economics and Statistics* 83: 663-674.

Dustmann, Christian and Arthur Van Soest (2002). “Language and the earnings of immigrants.” *Industrial and Labor Relations Review* 55: 473-492.

Dyen, Isidore, Joseph Kruskal and Paul Black (1992). “An Indo-European classification: A lexicostatistical experiment.” *Transactions of the American Philosophical Society* 82 (5).

[20] Of course, a spurt of teaching of English in school is well under way in China whereas the teaching of Chinese in English-speaking countries remains meager today. It would indeed be helpful to introduce school curricula in foreign languages in our model (with the appropriate lag) if it could be done (if the data was widely enough available). However, as emphasized earlier, it is not a foregone conclusion that major revision would follow: instruction in a foreign language as a child need not mean ability to converse in the language in adult life. The factors present in the model *may* still be the critical ones.

Egger, Peter and Farid Toubal (2015). “Languages and International Trade.” In *The Palgrave Handbook of Economics and Language.* Edited by Victor Ginsburgh and Shlomo Weber, in preparation.

Ethnologue. Available online at https://www.ethnologue.com.

Fearon, James (2003). “Ethnic and cultural diversity by country.” *Journal of Economic Growth* 8: 195-222.

Feely, Alan and Derek Winslow (2005). *Talking sense. A research study of language skills management in major companies*. London: CILT, The National Center for Languages.

Frankel, Jeffrey (1997). *Regional Trading Blocs in the World Trading System*. Washington DC: Institute for International Economics.

Frankel, Jeffrey and David Romer (1999). “Does trade cause growth?” *American Economic Review* 89: 379-399.

Fry, Richard and B. Lindsay Lowell (2003). “The value of bilingualism in the U.S. labor market.” *Industrial and Labor Relations Review* 57: 128-140.

Gabszewicz, Jean, Victor Ginsburgh and Shlomo Weber (2011). “Bilingualism and communicative benefits.” *Annals of Economics and Statistics* 101/102: 271-286.

Gaulier, Guillaume and Soledad Zignago (2010). BACI: International trade database at the product-level: The 1994-2007 version. CEPII Working Paper 2010-23.

Ginsburgh, Victor, Ignacio Ortuño-Ortin and Shlomo Weber (2007). “Learning foreign languages. Theoretical and empirical implications of the Selten and Pool model.” *Journal of Economic Behavior and Organization* 64: 337-347.

Ginsburgh, Victor and Juan Prieto (2011). “Returns to foreign languages of native workers in the European Union.” *Industrial and Labor Relations* 64: 599-618.

Ginsburgh, Victor and Shlomo Weber (2011). *How Many Languages Do we Need?* Princeton University Press.

Ginsburgh, Victor, Shlomo Weber and Sheila Weyers (2011) “The economics of literary translation: Some theory and evidence.” *Poetics* 39, 228-246.

Goldberger, Arnold (1964). *Econometric Theory*. New York: Wiley & Sons.

Graddol, David (1997). *The Future of English*. London: British Council.

Graddol, David (2006). *English Next*. London: British Council.

Grin, François (1999). *Compétences et récompenses: La valeur des langues en Suisse*. Fribourg: Éditions Universitaires de Fribourg.

Hagen, Stephen with James Foreman-Peck, Santiago Davila-Philippon, Bjorn Nordgren and Susanna Hagen (2006). *ELAN: Effects on the European economy of shortages of foreign language skills in enterprise*. Reading: CILT, The National Center for Languages.

Head, Keith, Thierry Mayer and John Ries (2010). “The erosion of colonial trade linkages after independence.” *Journal of International Economics* 81(1): 1-14.

Huntington, Samuel (1996). *The Clash of Civilizations and the Remaking of World Order*. New York: Simon & Schuster.

Kachru, Braj (2010). *Asian Englishes: Beyond the Canon*. Hong Kong University Press.

Irwin, Douglas and Marko Tervio (2002). “Does trade raise income? Evidence from the twentieth century,” *Journal of International Economics* 58: 1-18.

Laitin, David (1994). “The tower of Babel as a coordination Game: Political linguistics in Guana,” *American Political Science Review*, 88: 622-634.

Laitin, David (2000). “What is a language community?” *American Journal of Political Science* 44: 142-155.

Mayer, Thierry and Soledad Zignago (2011). “Notes on CEPII’s distances measures: the GeoDist Database.” CEPII Working Paper 2011-25.

Melitz, Jacques (2008). “Language and foreign trade.” *European Economic Review* 52: 667-699.

Melitz, Jacques (2015). “English as a global language.” In *The Palgrave Handbook of Economics and*

Language. Edited by Victor Ginsburgh and Shlomo Weber, in preparation.

Melitz, Jacques and Farid Toubal (2014). “Native language, spoken language, translation and foreign trade.” *Journal of International Economics* 93: 351-363.

Noguer, Marta and Marc Siscart (2005). “Trade raises income: A precise and robust result,” *Journal of International Economics* 65: 447-460.

Ostler, Nicholas (2010). *The Last Lingua Franca. English until the Return of Babel*. London: Allen Lane.

Penn World Tables. Available online at https://pwt.sas.upenn.edu.

Pool, Jonathan (1991). “The official language problem,” *American Political Science Review* 85: 495-514.

Pool, Jonathan (1996). “Optimal language regimes for the European Union,” *International Journal of the Sociology of Language* 121: 159-179.

Rodriguez, Francisco and Dani Rodrik (2001). “Trade policy and economic growth: A skeptic’s guide to cross-national evidence,” in Ben Bernanke and Ken Rogoff, eds. *NBER Macroeconomics Manual*, 261-325.

Selten, Reinhard and Jonathan Pool (1991). “The distribution of foreign language skills as a game equilibrium.” In *Game Equilibrium Models*, vol. 4. Edited by Reinhard Selten. Berlin: Springer-Verlag, 64-84.

Shy, Oz (2001). *The Economics of Network Industries*. Cambridge: Cambridge University Press.

Special Eurobarometer (2006). Europeans and their languages. *Special Eurobarometer* 243. Brussels: The European Commission.

Swadesh, Morris (1952). “Lexico-statistic dating of prehistoric ethnic contacts.” *Proceedings of the American Philosophical Society* 96: 121-137.

The British Chambers of Commerce (2003-2004). *BBC language survey. The impact of foreign languages on British business*. Part I, 2003, Part II, 2004. London: The British Chambers of Commerce.

Vaillancourt, François (1996). “Language and economic status in Quebec: Measurements, findings, determinants and policy costs.” *International Journal of the Sociology of Language* 121: 69-92.

Wooldridge, Jeffrey (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge; MA: The MIT Press.

Wooldridge, Jeffrey (2003). *Introductory Econometrics*. Mason, Ohio: Thomson-Southwestern.

Table 1. Destination languages (millions of speakers)

Language	Mother tongue	Worldwide speakers	Language multiplier
(1)	(2)	(3)	(4)=(3)/(2)
Arabic	244	272	1.11
Chinese	1161	1165	1.00
Dutch	22	37	1.68
English	357	1123	3.15
French	69	260	3.77
German	89	168	1.89
Italian	64	77	1.20
Japanese	126	126	1.00
Malay	22	158	7.18
Portuguese	209	222	1.06
Russian	184	267	1.45
Spanish	401	479	1.19
Turkish	91	102	1.12

Table 2: Foreign language learning

	Full sample			Positive sample		
	Probit	IV Probit		OLS	2SLS	
		First stage	Second stage		First stage	Second stage
	(1)	(2)	(3)	(4)	(5)	(6)
Speakers of acquired languages (log)	0.014***	0.006***	0.002***	0.024*	0.020***	0.008
	(4.238)	(3.565)	(2.823)	(1.833)	(2.683)	(0.408)
Speakers of native languages (log)	-0.016***	-0.001**	-0.003***	-0.025***	0.003	-0.027***
	(-4.335)	(-2.239)	(-4.307)	(-4.669)	(0.719)	(-4.328)
Trade with acquired language countries	0.440***		0.122***	0.803***		1.338***
	(8.977)		(3.076)	(4.638)		(3.057)
Distance between native and acquired language	-0.196***	-0.050***	-0.027***	-0.216**	0.030	-0.215**
	(-6.774)	(-5.423)	(-5.519)	(-2.366)	(0.808)	(-2.369)
Literacy rate in learning countries	0.215***	-0.012	0.025***	0.005	-0.114	0.068
	(4.892)	(-1.479)	(3.546)	(0.050)	(-1.644)	(0.553)
Instrument (FR native-language weighted)		0.570***			0.454***	
		(7.542)			(3.052)	
No. of observations	2,365	2,365	2,365	240	240	240
(pseudo) R-squared	0.242	0.146	–	0.237	0.157	0.172
No. of countries	193	193	193	94	94	94

Student *t*s in parentheses. These are based on robust standard errors clustered at country level.

*** p<0.01, ** p<0.05, * p<0.1. Intercepts are not reported.

Table 3: Foreign language learning with former colonial ties and Indo-European Dummy

	Full sample			Positive sample		
	Probit	IV Probit		OLS	2SLS	
		First stage	Second stage		First stage	Second stage
	(1)	(2)	(3)	(4)	(5)	(6)
Speakers of acquired languages (log)	0.016***	0.007***	0.001***	0.025*	0.017**	0.008
	(5.376)	(4.681)	(3.292)	(1.973)	(2.310)	(0.537)
Speakers of native languages (log)	-0.012***	0.000	-0.002***	-0.023***	0.003	-0.026***
	(-3.809)	(0.206)	(-3.376)	(-4.093)	(1.044)	(-4.269)
Trade with acquired language countries	0.263***		0.073***	0.637***		1.221***
	(6.804)		(2.607)	(3.863)		(3.766)
Distance between native and acquired language	-0.148***	-0.005	-0.020***	-0.281***	-0.010	-0.251***
	(-5.643)	(-0.462)	(-3.695)	(-3.059)	(-0.313)	(-2.794)
Literacy rate in learning countries	0.222***	-0.005	0.020***	0.114	0.005	0.113
	(5.581)	(-0.713)	(3.919)	(0.903)	(0.088)	(0.919)
Colonial language dummy	0.407***	0.116***	0.032**	0.149***	0.122***	0.075
	(8.693)	(5.553)	(2.206)	(2.888)	(5.683)	(1.218)
Indo-European dummy	0.044***	0.036***	0.004**	0.011	0.085***	-0.023
	(4.844)	(7.702)	(2.531)	(0.321)	(5.904)	(-0.553)
Instrument (FR native-language weighted)		0.539***			0.541***	
		(7.166)			(3.426)	
No. of observations	2,365	2,365	2,365	240	240	240
(pseudo) R-squared	0.304	0.229	–	0.276	0.346	0.211
No. of countries	193	193	193	94	94	94

Student *t*s in parentheses. These are based on robust standard errors clustered at country level.
*** p<0.01, ** p<0.05, * p<0.1. Intercepts are not reported.

.

Table 4: Foreign language learning without Chinese and Arabic

Panel A: Full Sample

	Without Chinese		Without Arabic		Without Chinese & Arabic	
	Probit	IV Probit Second stage	Probit	IV Probit Second stage	Probit	IV Probit Second stage
Speakers of acquired	0.032***	0.004***	0.014***	0.002***	0.035***	0.004***
languages (log)	(6.739)	(4.967)	(4.300)	(3.066)	(7.043)	(4.914)
Speakers of native	-0.017***	-0.003***	-0.017***	-0.003***	-0.017***	-0.003***
languages (log)	(-4.147)	(-4.093)	(-4.227)	(-4.283)	(-3.987)	(-3.999)
Trade with acquired	0.433***	0.115***	0.454***	0.110***	0.437***	0.104***
language countries	(8.172)	(2.820)	(8.725)	(3.277)	(7.784)	(2.994)
Distance between nat.	-0.178***	-0.024***	-0.194***	-0.026***	-0.165***	-0.022***
and acq. language	(-5.876)	(-4.635)	(-6.331)	(-5.437)	(-5.124)	(-4.228)
Literacy rate in learning	0.232***	0.026***	0.232***	0.029***	0.254***	0.030***
countries	(4.920)	(3.570)	(4.900)	(4.348)	(4.961)	(4.401)
No. of observations	2,176	2,176	2,193	2,193	2,004	2,004
(pseudo) R-squared	0.254	–	0.246	–	0.260	–
No. of countries	193	193	193	193	193	193

Panel B: Positive Sample

	Without Chinese		Without Arabic		Without Chinese & Arabic	
	OLS	2SLS Second stage	OLS	2SLS Second stage	OLS	2SLS Second stage
Speakers of acquired	0.029**	0.013	0.024*	0.007	0.029**	0.012
languages (log)	(2.256)	(0.674)	(1.838)	(0.408)	(2.271)	(0.676)
Speakers of native	-0.025***	-0.027***	-0.024***	-0.025***	-0.024***	-0.025***
languages (log)	(-4.600)	(-4.260)	(-4.226)	(-3.968)	(-4.154)	(-3.891)
Trade with acquired	0.803***	1.347***	0.807***	1.328***	0.807***	1.341***
language countries	(4.603)	(3.051)	(4.614)	(3.461)	(4.574)	(3.459)
Distance between nat.	-0.206**	-0.204**	-0.216**	-0.216**	-0.204**	-0.205**
and acq. language	(-2.248)	(-2.251)	(-2.285)	(-2.280)	(-2.152)	(-2.148)
Literacy rate in learning	0.008	0.071	0.080	0.152	0.082	0.157
countries	(0.074)	(0.579)	(0.845)	(1.317)	(0.871)	(1.347)
No. of observations	238	238	231	231	229	229
R-squared	0.241	0.174	0.238	0.175	0.242	0.176
No. of countries	94	94	93	93	93	93

Student *t*s in parentheses. These are based on robust standard errors clustered at country level.

*** p<0.01, ** p<0.05, * p<0.1. Intercepts are not reported.

Table 5: Foreign language learning with openness

	Full sample			Positive sample		
	Probit	IV Probit		OLS	2SLS	
		First Stage	Second Stage		First Stage	Second Stage
	(1)	(2)	(3)	(4)	(5)	(6)
Speaker of acquired languages (log)	0.014***	0.006***	0.002***	0.024*	0.020***	0.008
	(4.208)	(3.548)	(2.740)	(1.823)	(2.664)	(0.408)
Speaker of native languages (log)	-0.017***	-0.002**	-0.003***	-0.025***	0.003	-0.027***
	(-4.670)	(-2.403)	(-4.560)	(-4.587)	(0.741)	(-4.222)
Trade with acquired language countries	0.437***		0.123***	0.802***		1.339***
	(8.999)		(3.042)	(4.633)		(3.056)
Distance between native and	-0.197***	-0.051***	-0.026***	-0.216**	0.028	-0.214**
acq. language	(-6.843)	(-5.552)	(-5.481)	(-2.352)	(0.763)	(-2.350)
Literacy rate in learning countries	0.206***	-0.015*	0.023***	0.001	-0.120*	0.068
	(4.548)	(-1.936)	(3.078)	(0.012)	(-1.661)	(0.559)
Instrument (FR native-language weighted)		0.572***			0.453***	
		(7.577)			(3.039)	
Openness (log)	0.011	0.004***	0.002	0.003	0.005	-0.000
	(1.140)	(3.083)	(1.482)	(0.157)	(0.328)	(-0.023)
Observations	2,365	2,365	2,365	240	240	240
(Pseudo) R-squared	0.244	0.148	–	0.238	0.158	0.172
No. of countries	193	193	193	94	94	94

Student *t*s in parentheses. These are based on robust standard errors clustered at country level.
*** p<0.01, ** p<0.05, * p<0.1. Intercepts are not reported.

Table 6: Foreign language learning without large minority language

	Full sample				Positive sample	
	Probit	IV Probit		OLS	2SLS	
		First stage	Second stage		First stage	Second stage
	(1)	(2)	(3)	(4)	(5)	(6)
Speakers of acquired languages (log)	0.014***	0.005***	0.002***	0.030**	0.015**	0.025
	(4.359)	(3.076)	(4.049)	(2.232)	(2.036)	(1.453)
Speakers of native languages (log)	-0.016***	-0.001**	-0.002***	-0.023***	0.003	-0.024***
	(-4.334)	(-2.180)	(-4.159)	(-4.023)	(0.935)	(-3.876)
Trade with acquired language countries	0.412***		0.080***	0.793***		0.973**
	(8.880)		(2.770)	(4.342)		(2.555)
Distance between native and acquired language	-0.188***	-0.046***	-0.025***	-0.237**	0.052	-0.239**
	(-6.662)	(-5.040)	(-5.711)	(-2.471)	(1.351)	(-2.515)
Literacy rate in learning countries	0.204***	-0.013*	0.020***	-0.024	-0.136**	-0.001
	(4.693)	(-1.734)	(3.353)	(-0.215)	(-2.040)	(-0.010)
Instrument (FR native-language weighted)		0.625***			0.641***	
		(8.186)			(4.471)	
No. of observations	2,351	2,351	2,351	228	228	228
(Pseudo) R-squared	0.241	0.150	–	0.239	0.202	0.231
No. of countries	193	193	193	90	90	90

Student *t*s in parentheses. These are based on robust standard errors clustered at country level.
*** p<0.01, ** p<0.05, * p<0.1. Intercepts are not reported.

Table 7: Adding domestic language learning

	Full sample			Positive sample		
	Probit	IV Probit		OLS	2SLS	
		First stage	Second stage		First stage	Second stage
	(1)	(2)	(3)	(4)	(5)	(6)
Speakers of acquired languages (log)	0.018*** (4.565)	0.008*** (4.765)	0.004*** (3.513)	0.016 (1.180)	0.024*** (3.721)	0.034** (2.004)
Speakers of native languages (log)	-0.024*** (-4.754)	-0.001 (-1.289)	-0.005*** (-6.050)	-0.020*** (-3.313)	0.008** (2.238)	-0.016** (-2.275)
Trade with acquired language countries	0.480*** (7.858)		0.062* (1.799)	0.378*** (3.414)		-0.068 (-0.191)
Distance between native and acq. language	-0.324*** (-13.502)	-0.067*** (-6.277)	-0.053*** (-8.021)	-0.098** (-2.545)	0.014 (0.637)	-0.107*** (-2.805)
Literacy rate in learning countries	0.144*** (3.444)	-0.001 (-0.220)	0.007 (1.166)	-0.162** (-2.252)	0.012 (0.282)	-0.144* (-1.943)
Instrument (FR native-language weighted)		0.541*** (10.626)			0.415*** (5.748)	
No. of observations	2,502	2,502	2,502	345	345	345
(Pseudo) R-squared	0.286	0.233	–	0.082	0.216	0.031
No. of countries	193	193	193	158	158	158

Student *t*s in parentheses. These are based on robust standard errors clustered at country level.

*** p<0.01, ** p<0.05, * p<0.1. Intercepts are not reported.

Table 8: Adding country fixed effects

	Full sample			Positive sample		
	Probit	IV Probit		OLS	2SLS	
		First stage	Second stage		First stage	Second stage
	(1)	(2)	(3)	(4)	(5)	(6)
Speakers of acquired languages (log)	0.026*** (4.127)	0.004** (2.466)	0.002*** (2.840)	0.067*** (3.937)	-0.009 (-1.344)	0.070*** (4.378)
Speakers of native languages (log)	-	-	-	-	-	-
Trade with acquired language countries	1.560*** (7.537)		0.148** (2.519)	1.661*** (4.384)		2.206** (2.495)
Distance between native and acq. language	-0.449*** (-7.288)	-0.092*** (-4.958)	-0.017** (-2.549)	-0.135 (-0.765)	-0.077 (-1.487)	-0.083 (-0.557)
Literacy rate in learning countries	-	-	-	-	-	-
Instrument (FR native-language weighted)		0.557*** (5.661)			0.344*** (3.543)	
Country FE	Yes	Yes	Yes	Yes	Yes	Yes
No. of observations	1,173	1,173	1,173	240	240	240
(Pseudo) R-squared	0.487	0.221	–	0.623	0.751	0.603
No. of countries	94	94	94	94	94	94

Student *t*s in parentheses. These are based on robust standard errors clustered at country level.
*** p<0.01, ** p<0.05, * p<0.1. Intercepts are not reported. We lose observations in the full sample regressions because the dependent variable becomes perfectly predictable in some cases.

Table 9: Residuals of principal IV regressions by language

	Full sample			Positive sample		
Language	Mean[(a)]	Std. dev.	*t*-value	Mean[(a)]	Std. dev.	*t*-value
Arabic	-0.113	0.547	-0.207	0.039	0.211	0.185
Chinese	-0.212	0.363	-0.585	-0.197	0.016	-12.362
Dutch	-0.176	0.320	-0.550	-0.009	0.144	-0.066
English	-0.420	1.272	-0.331	0.081	0.292	0.276
French	0.081	0.713	0.114	0.038	0.190	0.199
German	-0.040	0.604	-0.066	-0.053	0.120	-0.440
Italian	-0.027	0.639	-0.043	-0.083	0.095	-0.879
Japanese[(b)]	-0.178	0.463	-0.386			
Malay	-0.047	0.237	-0.199	0.347	0.229	1.517
Portuguese[(b)]	-0.171	0.225	-0.760	-0.196		
Russian	0.088	0.565	0.157	0.019	0.202	0.095
Spanish	-0.126	0.992	-0.127	-0.090	0.102	-0.887
Turkish	-0.073	0.300	-0.245	0.010	0.117	0.089

(a) Estimates of the positive sample are based on Pearson residuals from the Probit regression in Table 2, column 3, and those of the positive sample are based on the IV regression in Table 2, column 6.

(b) Portuguese is acquired only in Spain (no standard deviation). Japanese is not acquired.

APPENDIX: Table A1: Summary Statistics

	Dimension	Mean	Standard Deviation
Full Sample (2365 observations)			
Foreign language learning	[0,1]	0.02	0.09
Speakers of acquired languages	Log	18.67	1.09
Speakers of native languages	Log	18.55	2.20
Trade with acquired language	[0,1]	0.05	0.09
Distance between native and acq. language	[0,1]	0.87	0.18
Literacy rate in learning countries	[0,1]	0.84	0.20
Colonial language dummy	(0,1)	0.02	0.16
Indo-European dummy	(0,1)	0.61	0.49
Openness	Log	-1.18	0.84
Positive Sample (240 observations)			
Foreign language learning	[0,1]	0.19	0.23
Speakers of acquired languages (log)	Log	18.94	0.80
Speakers of native languages	Log	17.28	2.04
Trade with acquired language	[0,1]	0.13	0.11
Distance between native and acq. language	[0,1]	0.77	0.20
Literacy rate in learning countries	[0,1]	0.93	0.12
Colonial language dummy	(0,1)	0.15	0.36
Indo-European dummy	(0,1)	0.93	0.25
Openness	Log	-1.04	0.69

APPENDIX: Table A2: Alternative one-part or single-equation estimates (marginal effects)

In Table A3, we present three alternative single-equation estimates of our basic model: IV Poisson, IV Negative Binomial, and IV Fractional Logit. IV Poisson is subject to the problem of overdispersion. Negative Binomial corrects for it. IV Fractional Logit also makes sense since the dependent variable is restricted to the unit interval [0, 1]. We tried Tobit as well. But the disturbances in the equation for the latent dependent variable fail all tests of normality and do so roundly. Therefore we do not present this last estimate. All three estimates in the Table are quasi maximum likelihood ones. We report the marginal effects at the means.

	IV Poisson	IV Negative Binomial	IV Fractional Logit Model
Speaker of acquired languages (log)	0.0024***	0.0023***	0.0021***
	(3.927)	(3.857)	(3.416)
Speaker of native languages (log)	-0.0020***	-0.0020***	-0.0020***
	(-5.610)	(-5.603)	(-5.528)
Trade with acquired language countries	0.0919***	0.0945***	0.1036***
	(5.975)	(5.938)	(5.377)
Distance between native and acquired language	-0.0382***	-0.0369***	-0.0370***
	(-6.400)	(-6.393)	(-6.072)
Literacy rate in learning countries	0.0290***	0.0271***	0.0258***
	(2.911)	(2.883)	(2.806)
Observations	2,365	2,365	2,365

Student *t*s in parentheses. These are based on robust standard errors clustered at country level. *** p<0.01, ** p<0.05, * p<0.1. Intercepts are not reported.

It is clear that all three estimates closely resemble one another and that the signs and significance of the influences correspond to those in Table 2. The numerical values are also close to those in the full sample estimate of the two-part model in Table 2. But we reject the idea of uniform behavior regardless of zero or positive learning.