DISCUSSION PAPER SERIES

No. 10101

FOREIGN LANGUAGE LEARNING: AN ECONOMETRIC ANALYSIS

Victor Ginsburgh, Jacques Mélitz
and Farid Toubal

INTERNATIONAL MACROECONOMICS and PUBLIC POLICY

Centre for Economic Policy Research

www.cepr.org

Available online at: **www.cepr.org/pubs/dps/DP10101.php**

ISSN 0265-8003

FOREIGN LANGUAGE LEARNING: AN ECONOMETRIC ANALYSIS

Victor Ginsburgh, Université libre de Bruxelles, CORE, Université catholique de Louvain, and Ural Federal University
Jacques Mélitz, ENSAE, CEPII, and CEPR
Farid Toubal, Ecole Normale Supérieure de Cachan, Paris School of Economics and CEPII

Discussion Paper No. 10101
August 2014

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801, Fax: (44 20) 7183 8820
Email: cepr@cepr.org, Website: www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programme in **INTERNATIONAL MACROECONOMICS and PUBLIC POLICY**. Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

CEPR Discussion Paper No. 10101

August 2014

ABSTRACT

Foreign Language Learning: An Econometric Analysis

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. Four factors significantly explain learning, two of which affect the broad decision to learn, while two concern as well the choice of the particular language to learn. Literacy generally promotes learning while the world population of speakers of the native language generally discourages it. Trade with speakers of a specific language prompts learning of that specific language while the linguistic distance between the home and the foreign language discourages learning of the specific language. Trade is highly significant and may well deserve more emphasis than the other three key variables (literacy rate, linguistic distance, and world population of native speakers) because its direction can change faster and by a larger order of magnitude. Controlling for individual acquired languages, including English, is of no particular importance.

JEL Classification: F10, F20 and Z00
Keywords: English as a global language, language and trade and language learning

Victor Ginsburgh
ECARES
Universite Libre de Bruxelles
CP 114
Avenue F Roosevelt 50
1050 Bruxelles
BELGIUM

Email: vginsbur@ulb.ac.be

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=101991

Jacques Mélitz
Département de la Recherche
CREST-INSEE
15 Bd. Gabriel Péri
92245 Malakoff CEDEX
FRANCE

Email: j.melitz@hw.ac.uk

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=100052

Farid Toubal
Ecole Normale Supérieure de Cachan
CES-Cachan – 61
avenue du Président Wilson
94235 Cachan
FRANCE

Email: ftoubal@ens-cachan.fr

Submitted 25 July 2014

1. Introduction

Multilingualism receives considerable attention by linguists, historians, philosophers, social scientists and literary writers and critics. Yet econometric work on language learning has lagged behind. It 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 EU. In our effort to pursue the same line of research as theirs, we explicitly consider only the learning of foreign languages and not that of the primary language in the home country, which we assume to be the dominant choice for daily living. To simplify our analysis, we extend this assumption to linguistic minorities, possibly concentrated in certain regions, like Basque speakers in Spain or Gujarati speakers in India. In addition, we take a world view of the subject and deal 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 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. Evidently, citizens of a country who are engaged in trade with different parts of the world may take different decisions about which language to learn.

The trade motive for language learning emerges as the most important factor in our empirical findings. 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 2.7 percentage points. This is a large effect. It emerges after controlling for the reciprocal effect of learning on the trade share,

*The authors would like to thank Olivier Loisel and the participants in seminars at CREST, Paris, and Ekaterinburg, Russia, for valuable 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.

[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).

that is, after instrumenting the trade share. In the absence of control for endogeneity, the effect is much smaller (about one-third as high). 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 26% probability of some positive learning where there is none.

But there are other factors as well. A large world population of speakers of the home language discourages learning, just as theory would say, while a high literacy rate does the opposite. Linguistic distances also have the expected effect. When the distance between languages increases, learning decreases. Interestingly, one of the predicted effects of theory does not emerge: The size of the world population of speakers of a foreign language does not encourage more learning. But while this effect shows up even after introducing the trade share, it becomes insignificant after instrumentation. It seems therefore that once the commercial incentives to learn a language are properly accounted for, one can no longer detect the non-commercial incentives to learn it, even though these refer to important matters such as ease of social interaction with people from different cultures, the benefits of access to their cultures and their literary and artistic heritages.

The paper proceeds as follows. Section 2 will provide the theoretical background for 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

The main tradition thus far in economic theorizing on language learning is to use game theory. Selten and Pool (1991) were first to publish a general game-theoretic model of language acquisition in which the payoff of each citizen-player in community K is a function of his own strategy and of the strategies of all players of other communities J, with the exception of his own. Learners of a foreign language incur a cost that is different for each individual. The authors show that under certain assumptions, there exists a language acquisition Nash equilibrium. Church and King (1993) and Shy (2001) follow similar steps in a simplified model. They study a situation with two groups in the same bilingual country or in two different countries, and two native languages. Citizens know their regional or national language, but consider acquiring the other language at a cost that is identical for all citizens in each region or country. The benefit to a citizen increases with the number of people with whom he can communicate in a common language, whether in his native language, or in the other one. The Nash equilibrium results in corner equilibria only: Either no one learns the foreign language in either country (if the payoff of learning is sufficiently low), or everybody learns the foreign language in one country while nobody does in the other. This results from

the fact that, though payoffs vary across countries, they are identical across citizens in each country. Gabszewicz, Ginsburgh and Weber (2011) (GGW in what follows) take these models a step forward by going back to Selten and Pool (1991) and recognizing heterogeneous populations. Thereby they show that interior equilibria as well as corner equilibria can exist. Like Selten and Pool, GGW base the heterogeneity of citizens strictly on differences in their aptitude to learn the foreign language.

In this work, we use similar reasoning to GGW to explain the learning decision but we drop the game theoretical aspect. Each individual rests his decision to learn on the current situation he faces independently of what others are doing. The others' current decisions are simply too small to affect him. Stated differently, the conditions in his environment and his abilities dominate his decision, regardless of the current learning decisions of others. We shall also broaden the basis for the heterogeneity of learning decisions to encompass differences in individual returns to learning and the opportunity cost of time spent on learning as well as differences in aptitudes to learn.

Let there be M countries and L languages where $L \leq M$.[4] Each individual is a resident of a single country and knows at least one language (usually of the country of birth, but not necessarily so). There is a primary language in each country whose learning is ignored. We thus disregard the decision of a German resident to learn German in Germany, but do, however, consider the decision of German residents to learn Turkish and American residents to learn Spanish though there are native Turkish-speakers in Germany and native Spanish-speakers in the US. We will return later to the exact line of demarcation, which is of course important.

With more than two languages, the time spent on learning one language is at the expense of learning another. This is a vital aspect of heterogeneity. Not only may some people choose to learn while others choose not to, but some people may chose to learn one language and others to learn another. We shall assume that it is too costly to learn more than one language. Therefore, there is major competition between languages.

Consider then the individual in country K who already possesses language K (not everybody) but wishes to decide whether to learn foreign language J with $J = 1, 2, \ldots, K-1, K+1, \ldots, L$. In line with GGW's discussion and the tradition they follow, suppose that the individual's decision depends on the additional number of people with whom he will be able to communicate if he learns the language and on the cost of learning. Let the world population that knows language K be N_K and the world population that knows language J be N_J. Let $B(N_K)$ be the benefit of an individual living in country K from his current language repertoire

[4] Of course, in reality there exist many more languages (over 6,000) than countries (around 200).

which we assume for the moment to consist of his primary language *K* only. Assume also that $B'(N_K) > 0$ *and* $B''(N_K) < 0$.

If the individual learns language *J* his benefit will be $B(N_J + N_K)$. Therefore, his payoff from learning language *J* is

$$P_{JK} = B(N_J + N_K) - B(N_K) - C_{JK}(.), \text{ for all } J \neq K, \quad (1)$$

with $\partial P_{JK}/\partial N_J > 0$, sign $\partial P_{JK}/\partial N_K$ = sign $B''(N_J + N_K) < 0$, and $\partial P_{JK}/\partial C_{JK} = -1$.

C_{JK} refers to the cost of learning, a function of several variables. Let us assume that:

$$C_{JK} = C_{JK}(tr_{JK}, D_{JK}, R_K) \quad (2)$$

where tr_{JK} is the individual's total trade with the *J*-speaking world (both as buyer and seller), with $C'_{JK}(tr_{JK}, . , .) < 0$; D_{JK} is the linguistic distance between languages *J* and *K* and $C'_{JK}(. , . , D_{JK}) > 0;$ and R_K is equal to 1 or 0 depending on whether the person is literate with $C_{JK}(. , . , 1) < C_{JK}(. , . , 0)$. Thus, trade with the *J*-speaking world reduces the opportunity cost of time spent learning *J*; linguistic distance between *J* and *K* makes language *J* more difficult to learn; and literacy helps learning.

Each individual compares the $L - 1$ payoffs. If all of them are 0 or negative, he learns no language. If one or more are positive, he learns the one with the highest payoff. We assume that it is too costly to learn more.

All individuals in country *K* face the same decision problem. However, payoffs and therefore decisions will differ for many reasons. Different cultural interests lead to different responses to N_J. Different lines of economic activity and tastes mean different individual values of tr_{JK} and people may also respond differently to the same value. In addition, differences in aptitudes to learn may cause different responses to linguistic distances D_{JK}. Finally, literacy R_K is not the same for everyone. As a result, individuals will learn different languages; some people will not learn at all; and there may also (and will probably be) no learning of some languages. For convenience, we shall assume some learning of some language in every country. Otherwise a uniform treatment everywhere would be less warranted.

Let α_{JK} be the share of the population in country *K* that learns language *J*. Based on the previous analysis of individual behavior, α_{JK} may be written:

$$\alpha_{JK} = \text{F } (N_J, N_K, T_{JK}, D_{JK}, I_K), \text{ for all } J \text{ and } K, J, J \neq K \quad (3)$$

where N_J, N_K and D_{JK} are as before, but tr_{JK} and R_K are replaced since they differ across

individuals. The ratio of the total trade of country *K* with the *J*-speaking world, T_{JK}, serves instead of tr_{JK}, and the literacy rate in country *K*, I_K, serves instead of R_K. Quite specifically, as regards T_{JK}:

$$T_{JK} = \frac{\Sigma_{h \in H} \sigma_{hJ} T_{hK}}{\Sigma_{h \in H} T_{hK}} \tag{4}$$

where *H* is the set of country *K*'s trading partners, σ_{hJ} is the share of native speakers of language *J* in country *h*, T_{hK} is the total trade of country *K* with country *h*, and the denominator in eq. (4) is therefore country *K*'s total trade.

The earlier theoretical signs in eqs. (1) and (2) imply that α_{JK}, the share of the total population in country *K* that learns language *J*, is:

(i) increasing in N_J: the larger the number of speakers of the acquired (or destination) language *J*, the more *J* is attractive;
(ii) decreasing in N_K: the larger the number of speakers of the source language *K*, the less learning of any other language is needed;
(iii) increasing in T_{JK}: the greater the intensity of trade with foreign speakers of language *J*, the more learning of *J* is attractive;
(iv) decreasing in D_{JK}: the larger the distance between the source and the destination language, the less learning of *J* is attractive;
(v) and increasing in I_K: the more educated the home population, the more they will learn.

Two points deserve comment. We took as a basis for reasoning monolinguals but bilinguals in country *K* can communicate with more people than world speakers of language *K* without learning any new language. Therefore, if there are bilinguals, N_K in eq. (3) is a minimum of the total world population with whom the *K*-speakers in country *K* can already communicate and may thus be interpreted as a reflection of this larger world total. Next, T_{JK} may largely reflect the size of the world language community N_J. Therefore, in the presence of T_{JK}, N_J may possibly best be seen as a reflection of the non-market advantages of learning: that is, the ability to interact socially with native speakers of foreign languages and to benefit from their cultures and cultural heritages.

3. Econometric specification

We test a linear world approximation to eq. (1) consisting of $(L-g) \times M$ values[5] of α_{JK} as a function of the five right-hand side variables, understood to be exogenous. This exogeneity

[5] $g = 0$ for countries whose native language does not belong to the 13 languages that we study and $g = 1$ for those whose native language we do study (since they are left with the choice of learning one of 12 languages). However, for reasons that will be explained later, *g* can also equal 2.

can reasonably be accepted for linguistic distances and literacy rates. It also follows for N_J and N_K if we measure them on the basis of world native-language populations rather than total world speakers, as we will. However, though as already mentioned, we also base T_{JK} strictly on native speakers, the assumption of exogeneity is certainly not true for trade, since knowing a language promotes trade with native speakers as well as the rest. To deal with the endogeneity of T_{JK}, we instrument it by Y_{JK}, the share of the total domestic output (measured as GDP) of all of *K*'s trade partners attributable to native speakers of language *J*. Quite specifically,

$$Y_{JK} = \frac{\sum_{h \in H} \sigma_{hJ} GDP_{hK}}{\sum_{h \in H} GDP_{hK}} \tag{5}$$

where, as before in eq. (4), *H* is the set of *K*'s trading partners and σ_{hJ} is the share of native speakers of language *J* in country *h*. The foreign output share Y_{JK} is expected to be positively related to T_{JK} (despite the notable absence of trade weights in eq. (5)) whereas except for an effect on *H*, which we consider negligible,[6] any learning of language *J* in country *K* should have a negligible effect on Y_{JK}.[7]

The equation to estimate reads:

$$\alpha_{JK} = \beta_0 + \beta_1 N_J + \beta_2 N_K + \beta_3 T_{JK} + \beta_4 D_{JK} + \beta_5 I_K + \varepsilon_{JK} \tag{6}$$

where T_{JK} is instrumented by (5). 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.[8] 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.

Two additional control variables need consideration. One is a dummy variable C_K 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 *IE* for Indo-European languages. Among the 13 destination languages in our study,

[6] Note that trade is widespread between countries with no native speakers of a common language on the basis of third languages or translators and interpreters. In all these cases, as well as those of trade between countries with the same native language, learning does not affect *H*.

[7] We also experimented with a second instrument for T_{JK}: the geographical distance of country *K* to speakers of language *J*, as measured by the sum of the distances of country *K* to all other countries with weights for the individual distances depending on the percentages of native speakers of language *J* in the respective foreign countries. There was no improvement: Y_{hK} alone does as well.

[8] 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.

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 (at least, if both belong to the same family, 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 *IE* has a positive effect.

There are two basic reasons for numerous $\alpha_{JK} = 0$ values, one of which, the first, has already been mentioned. Each individual learns a small number of languages at best, in fact at most only one, according to our modeling. Many zeros will appear on this count, even at the national level. Second, the number of learners of any particular language may be small and we only collect values of α_{JK} that are at least equal to one percent at the national level. Thus, despite our assumption of some learning of some foreign language in every country, the positive values may be too small to appear. This last assumption of positive learning everywhere is useful since, however small the number of learners of a foreign language may be, in a maximizing framework the factors determining their behavior will determine total learning in the country.

4. Data

The necessary data requires 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 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 used Wikipedia, website http://en.wikipedia.org/wiki/List_of _countries_by_English-speaking_population (downloaded 18 June 2010), which reproduces the same numbers that we had extracted from the EU survey but also updates some of the estimates in Crystal *(*2003a, p. 109) for the rest of the world on the basis of various national census reports and more recent sources. 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).

– For French, we used the "Estimation du nombre de francophones dans le monde" website http://www.axl.cefan.ulaval.ca/francophonie/OIF-francophones-est2005.htm, completed by information from the separate Wikipedia websites for "African French" (http://en.wikipedia.org/wiki/African_French) and for "French Language" (http://en. wikipedia.org/wiki/French_language); all the figures come from referenced French governmental sources.

– For German, we used Wikipedia's website http://en.wikipedia.org/wiki/German _as_a_minority_language (referenced sources).

– For Spanish, we used Wikipedia's website http://en.wikipedia.org/wiki/Spanish _language (referenced sources).

– 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. 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.

The dependent variable in our model, α_{JK}, is the ratio of non-native speakers of language *J* in country *K* and the number of inhabitants of country *K*. The N_J values follow directly from the world values of native speakers in levels. There are 13 N_J values. On the other hand, N_K varies 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.

We also faced the problem of choosing a primary language for each country, not only to decide which learning decisions to drop but also to define 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 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,

[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 1 times 0.6 plus worldwide native speakers of language 2 times 0.4. In fact, the percentage values for native languages in the database usually add up to less than one, sometimes much less, as any attempt to avoid this would have meant adding hundreds, if not thousands, more languages in the analysis. The sums less than one also lead to lower N_K figures. However, this is of little importance since the omitted contributions to N_K are generally small, all the more so after applying the national weights to the world figures (because the languages themselves are small or because the weights are small or both). All in all, 106 different languages enter in the determination of the N_K values for all 193 countries. N_K is never zero since we always include the largest language in a country.

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, 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 are represented by 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 will be represented by 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]

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). Note that we depart from the recent practice, stemming from Laitin (2000) and Fearon (2003), of founding the linguistic distances on the *Ethnologue* classification of language trees.[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 multiply K's bilateral trade with each of its trade partners by the respective percentage of native speakers of language J in the partner country, sum over all partner countries and divide by the total trade of country K (see eq. (4)). 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 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]

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}. The zero values reflect cases where learning in a country K is dominated by other languages (possibly but not necessarily one of our 13 destination languages) or where positive learning is below one percent. Significantly too, in the instances of domination by another language, there are some near hits and some wild misses. The 240 positive values seem more comparable with one another since they are all concerned with choices of learning. It does not appear reasonable to suppose that a single mechanism determines the numerous zeros that are

[10] See also http://wwwstaff.eva.mpg.de/~wichmann/ASJPHomePage.htm
[11] For example, for India, we weigh Hindi .67 and English .33.
[12] Melitz and Toubal (2014) experimented with both the ASJP measure of D_{JK} and the Fearon-Laitin one in a study of bilateral trade. Results were similar.
[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.

associated with a decision to learn (because of competition between languages, and even though heterogeneity will often reduce this number at the national level) and the wide array of positive values. Therefore we made two separate estimates of the basic model. First, we considered the binary choice between learning and not learning for the full sample and estimated the model using probit. Next, we considered the percentage of learners conditional on positive learning (240 observations) and applied ordinary least squares. In both cases we instrumented for trade, therefore using probit with instrumentation in the former and two stage least squares in the latter.[14]

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. (Table A1 of the Appendix provides summary statistics for our main 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 except the one for speakers of acquired languages whose sign becomes negative but not statistically different from zero. Based on the estimates, the largest effect by far on learning appears to be trade. Specifically, there is a 26% 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 (0.36) is not really strikingly higher, if at all, than the other three 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 (0.4). Yet trade is also more variable than these other three factors, especially linguistic distance, which is a constant, and the literacy rate, which is, in many cases, close to one. Thus, the emphasis on trade remains perhaps right.

Columns (4) to (6) (positive sample) deal with the results conditional on positive learning. Once again the instrument for trade performs well (column 5), but now the correction for endogeneity markedly raises the coefficient for trade at the margin. A one percentage-point increase in the ratio of trade with native speakers of the destination language would increase learning of the language by 2.66 percentage points, conditional on positive learning. The

[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.

effects of world population of native speakers, linguistic distance and literacy still come out with the right signs as before in the full sample, but the last two coefficients are significantly different from zero only at the 12% probability level (and in this case it does not much matter if we look at the standardized "beta" values of the coefficients instead as both of them are lower). The negative significant effect of native language on learning is of some consequence. A 100 percent increase of speakers of the native language would reduce learners of other languages by 2.9 percent. Thus, in a nation of 50 million native speakers in which there are already learners, this would mean a reduction of 1.45 million learners.

7. Robustness checks

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 changes little, we simply show the results of adding both jointly. As seen in Table 3, former colonial languages are highly important in both samples prior to correction for the endogeneity of trade. The same is true for Indo-European languages but only for the full sample, that is, for the question of the existence of learners and not how numerous they are. However, following instrumentation of trade, the colonial dummy ceases to be important in both samples. The Indo-European dummy, on its part, also behaves more poorly. It remains important in the full sample, though with a much lower coefficient. However, it even assumes the wrong negative sign in the positive-value sample, and significantly so at the 90% confidence level. The degradation of the results goes beyond the behavior of these two controls in this last sample. Both linguistic distance and literacy, which were marginally significant before slightly below conventional levels, are now clearly insignificant. The baseline model therefore seems satisfactory.[15]

The next robustness test simply reinforces this last conclusion. Since eight of the 13 destination languages are Indo-European, we can study this group separately rather than with an Indo-European dummy. Accordingly, we also ran the two types of regressions (full and positive-value samples) without the observations for non Indo-European source or destination languages. This meant a drop in the number of observations to 1,431 in the full sample and to 224 (proportionately much less) in the positive sample. Our results (not shown) hardly change. Thus, Indo-European languages do not behave differently than the rest.

[15] The poor performance of the colonial variable is clearly linked to the highly significant positive coefficient of the colonial language variable in both first stage equations for trade, that is, in columns 2 and 5 of Table 3. To all evidence, it is difficult to disentangle the effects of colonial language and trade. Indeed the results of Table 3 would even say that colonial language would be a fitting instrument for trade, alongside the foreign GDP ratios, since the variable significantly affects learning but essentially via trade. We did not proceed in this direction.

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 websites of the British Council and of Wikipedia. In fact, we predominantly repeat the same figures for individual countries that Crystal (2003b, p. 60) provides, which cover only 75 "territories where English has held and continues to hold a special place," 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 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 last three robustness checks are concerned with more conceptual issues.

First, though our trade variable focuses on relative trade in different languages, we consider that it reflects not only the desire to learn one particular language but also the desire to learn foreign languages in general. Notwithstanding, one could wonder whether the variable

adequately reflects the common influence of trade on the incentive to learn languages. Accordingly, 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 the product of the variable and world population of the destination language N_J, or else still the two simultaneously (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 0 and its presence hardly alters the other coefficients. The result is always the same regardless of 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. This takes care of both examples, that is, both of them drop out on grounds that domestic conditions rule in deciding whether to learn the language. Another 12 observations drop out as well (for the same reason).[16] As can be verified in Table 6, the loss of these 14 observations has almost no effect.

The third and last robustness check responds to a diametrically opposite question to the preceding 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) concerning learning in the full sample. These reflect the instances of no learning of our 13 languages even though they are primary. Results are shown in Table 7. The quality of the fit drops significantly, especially after instrumenting trade. If we compare these new results after instrumenting trade with those in Table 2 (columns 3 and 6), we find that the world population of speakers of the destination language, formerly not

[16] The 14 observations (including the two for Belgium and Latvia) 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 level of native speakers) or Turkish in Cyprus (despite the 20 percent level 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).

significantly different from zero, now becomes significantly negative for both samples, contrary to theory. The impact of linguistic distance drops nearly by half in the full sample regression and becomes insignificant in the other. Literacy, which was previously highly significant in the full sample, now becomes insignificant, and acquires the wrong, highly significant, negative sign in the positive-value sample. We conclude that the additional observations cannot be properly accounted for in our analysis: the decision to learn the primary language of a country by immigrants and other permanent residents is indeed a subject requiring separate analysis, since the incentives to learn are different.

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. Is this right? 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.

As an alternative, therefore, what we show in Table 8 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 8 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 a much larger standard deviation.

[18] The other positive values for Portuguese in our sample are for countries where the language is a primary one. Therefore we do not include these other cases.

The general impression from Table 8 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 –0.647 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 there should be none (accordingly the *t*-statistic is low, 0.41). Furthermore, in the restricted or positive-value sample, the mean error for English is almost zero and the lowest of all, which goes entirely against the idea of status as a *lingua franca*.[19] The case of Japanese deserves special mention too since there is no observation with positive learning for this language. Yet its mean residual of 0.21 with a *t*-statistic of 0.55 fits in well with the figures in the rest in the sample.

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 also noteworthy on 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 stay within the tradition of Selten and Pool (1991) and Church and King (1993) as well as the extension of the second paper by Gabszewicz, Ginsburgh and Weber (2011), except for the fact that we drop the current interactions between learners. Our only other notable modification is to add trade as a factor. This factor is important both in contributing to the heterogeneity of learning decisions and in reflecting the commercial inducements to learn foreign languages, the common as well as the idiosyncratic ones.

Our results, based on world data, support the view that a unified approach to language learning without any attention to particular languages has some merit. International trade has

[19] This is not to question that English is or might be a *lingua franca* in some limited areas like air traffic control, scientific writing and international sports. 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 (full sample) or the other (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.

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. This clearly agrees with general first-hand experience of foreign visitors to English-speaking countries such as Great Britain and the United States. Linguistic distances have a negative effect on learning while the effect of literacy is positive. But both effects are notably clearer for the decision to learn than for additional learning if there is some.

The hypothesis that a large population of world speakers attracts learning does not seem to hold, once proper account is given to trade by instrumentation. This could partly result from a problem of statistical inference. We only consider 13 destination languages, while we have many times more observations of the other major influences we investigate. 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 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 to learn fewer foreign languages while speakers of other languages will not wish to learn either more or less Arabic or Spanish. Thus, Arabic and Spanish will become relatively more important, as Graddol (2006) foresees. In theory, of course, these demographic assumptions would mean more learning of Arabic and Spanish in absolute terms, which would therefore reinforce the rise in the relative size of those two

[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 retarded 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, 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.

languages. According to our results, this reinforcing effect depends entirely on a rise in the share of Arabic and Spanish trade in non-Arabic and non-Spanish-speaking countries. Therefore, the reinforcing effect may not materialize. But in any event, 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: An Indo-European classification: A lexicostatistical experiment." *Transactions of the American Philosophical Society* 82 (5).

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.

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

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.

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.

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.

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

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

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

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

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.

Laitin, David (2000). "What is a language community?." *American Journal of Political*

Science 44: 142-155.

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

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

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.

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, 64-84. Berlin: Springer-Verlag.

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

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 (1)	Mother tongue (2)	Worldwide speakers (3)	Language multiplier (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***	0.001	-0.001	0.024*	-0.001	-0.032
	(4.348)	(0.720)	(-1.109)	(1.841)	(-0.154)	(-1.306)
Speakers of native languages (log)	-0.015***	-0.000	-0.003***	-0.024***	0.002	-0.029***
	(-3.992)	(-0.720)	(-4.049)	(-4.412)	(0.754)	(-3.384)
Trade with acquired language countries	0.465***		0.263***	0.788***		2.665***
	(9.243)		(3.828)	(4.688)		(4.129)
Distance between native and acquired language	-0.317***	-0.079***	-0.058***	-0.355**	-0.062	-0.279
	(-6.966)	(-4.657)	(-5.293)	(-2.197)	(-1.295)	(-1.633)
Literacy rate in learning countries	0.249***	0.010	0.041***	0.064	-0.109*	0.286
	(5.323)	(1.466)	(3.292)	(0.570)	(-1.852)	(1.536)
Instrument (GDP ratio)		0.524***			0.373***	
		(11.570)			(4.232)	
No. of observations	2,365	2,365	2,365	240	240	240
(pseudo) R-squared	0.234	0.202		0.236	0.156	
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***	0.003*	-0.000	0.024*	0.000	-0.044
	(5.265)	(1.972)	(-0.444)	(1.913)	(0.050)	(-1.423)
Speakers of native languages (log)	-0.011***	0.000	-0.002***	-0.021***	0.004	-0.033***
	(-3.484)	(0.268)	(-2.863)	(-3.772)	(1.369)	(-3.464)
Trade with acquired language countries	0.249***		0.156***	0.647***		3.161***
	(7.391)		(2.622)	(4.098)		(3.437)
Distance between native and acquired language	-0.182***	-0.068***	-0.042***	-0.441***	-0.117**	-0.196
	(-5.763)	(-3.933)	(-3.396)	(-2.644)	(-2.525)	(-0.959)
Literacy rate in learning countries	0.211***	0.013*	0.032***	0.189	-0.007	0.229
	(5.740)	(1.871)	(3.316)	(1.494)	(-0.147)	(1.257)
Colonial language dummy	0.301***	0.090***	0.012	0.138***	0.106***	-0.149
	(7.356)	(3.930)	(1.208)	(2.727)	(4.606)	(-1.046)
Indo-European dummy	0.071***	0.010***	0.007**	-0.025	0.003	-0.118*
	(6.526)	(2.645)	(2.486)	(-0.443)	(0.152)	(-1.648)
Instrument (GDP ratio)		0.456***			0.313***	
		(9.716)			(3.519)	
No. of observations	2,365	2,365	2,365	240	240	240
(pseudo) R-squared	0.322	0.230		0.270	0.250	
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

	Full sample						Positive sample					
	Without Chinese		Without Arabic		Without Chinese & Arabic		Without Chinese		Without Arabic		Without Chinese & Arabic	
	Probit	IV Probit Second stage	Probit	IV Probit Second stage	Probit	IV Probit Second stage	OLS	TSLS Second stage	OLS	TSLS Second stage	OLS	TSLS Second stage
Speakers of acquired languages (log)	0.034***	0.001	0.015***	-0.002	0.038***	0.001	0.029**	-0.022	0.025*	-0.045	0.030**	-0.034
	(7.065)	(0.870)	(4.569)	(-1.256)	(7.484)	(0.917)	(2.265)	(-0.922)	(1.860)	(-1.590)	(2.298)	(-1.242)
Speakers of native languages (log)	-0.016***	-0.003***	-0.016***	-0.003***	-0.016***	-0.003***	-0.023***	-0.028***	-0.022***	-0.026***	-0.021***	-0.025***
	(-3.855)	(-3.941)	(-3.903)	(-3.782)	(-3.741)	(-3.624)	(-4.358)	(-3.447)	(-4.028)	(-2.606)	(-3.970)	(-2.671)
Trade with acquired language countries	0.449***	0.221***	0.471***	0.272***	0.442***	0.217***	0.790***	2.515***	0.792***	2.982***	0.792***	2.808***
	(8.375)	(3.635)	(8.879)	(3.623)	(7.852)	(3.396)	(4.647)	(4.176)	(4.661)	(3.915)	(4.616)	(3.993)
Distance between nat. and acq. language	-0.299***	-0.053***	-0.311***	-0.056***	-0.281***	-0.049***	-0.340**	-0.269	-0.350**	-0.262	-0.334**	-0.253
	(-6.373)	(-5.020)	(-6.724)	(-4.854)	(-5.919)	(-4.497)	(-2.104)	(-1.605)	(-2.152)	(-1.463)	(-2.055)	(-1.451)
Literacy rate in learning countries	0.261***	0.038***	0.268***	0.053***	0.283***	0.048***	0.064	0.268	0.144	0.460**	0.143	0.434**
	(5.344)	(3.281)	(5.330)	(4.352)	(5.372)	(4.381)	(0.570)	(1.515)	(1.475)	(2.105)	(1.473)	(2.129)
No. of observations	2,176	2,176	2,193	2,193	2,004	2,004	238	238	231	231	229	229
(pseudo) R-squared	0.249		0.239		0.258		0.237		0239		0.243	
No. of countries	193	193	193	193	193	193	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. 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***	0.001	-0.001	0.024*	-0.002	-0.032
	(4.324)	(0.721)	(-1.130)	(1.838)	(-0.165)	(-1.297)
Speaker of native languages (log)	-0.016***	-0.000	-0.004***	-0.024***	0.003	-0.030***
	(-4.247)	(-0.798)	(-4.311)	(-4.321)	(0.817)	(-3.239)
Trade with acquired language countries	0.463***		0.260***	0.789***		2.666***
	(9.259)		(3.788)	(4.693)		(4.137)
Distance between native and acquired language	-0.317***	-0.080***	-0.057***	-0.355**	-0.062	-0.279
	(-6.876)	(-4.628)	(-5.255)	(-2.192)	(-1.295)	(-1.632)
Literacy rate in learning countries	0.242***	0.009	0.038***	0.066	-0.117*	0.304*
	(5.036)	(1.247)	(3.118)	(0.601)	(-1.850)	(1.645)
Openness (log)	0.008	0.002	0.002	-0.001	0.006	-0.013
	(0.882)	(1.369)	(0.838)	(-0.083)	(0.430)	(-0.375)
Observations	2,365	2,365	2,365	240	240	240
(Pseudo) R-squared	0.235	0.202		0.236	0.158	
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	
	(1)	First stage (2)	Second stage (3)	(4)	First stage (5)	Second stage (6)
Speakers of acquired languages (log)	0.014*** (4.457)	0.001 (0.460)	-0.002 (-1.459)	0.030** (2.208)	-0.008 (-0.816)	-0.024 (-1.007)
Speakers of native languages (log)	-0.015*** (-3.998)	-0.000 (-0.521)	-0.004*** (-4.024)	-0.022*** (-3.844)	0.002 (0.719)	-0.027*** (-2.958)
Trade with acquired language countries	0.435*** (9.137)		0.289*** (3.783)	0.773*** (4.391)		2.730*** (4.718)
Distance between native and acquired language	-0.307*** (-6.943)	-0.075*** (-4.562)	-0.060*** (-5.200)	-0.389** (-2.308)	-0.049 (-1.123)	-0.340* (-1.933)
Literacy rate in learning countries	0.236*** (5.118)	0.008 (1.192)	0.037*** (2.872)	0.037 (0.313)	-0.125** (-2.123)	0.302 (1.536)
Instrument (GDP ratio)		0.527*** (11.681)			0.413*** (5.098)	
No. of observations	2,351	2,351	2,351	228	228	228
(Pseudo) R-squared	0.233	0.205		0.238	0.177	
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)	0.015***	0.003*	-0.006***	0.015	0.012	-0.080***
	(3.781)	(1.956)	(-2.818)	(1.119)	(1.552)	(-2.696)
Speakers of native languages (log)	-0.022***	-0.000	-0.006***	-0.019***	0.004	-0.034***
	(-4.234)	(-0.454)	(-4.977)	(-3.299)	(1.466)	(-3.805)
Trade with acquired language countries	0.513***		0.432***	0.372***		2.734***
	(7.777)		(4.740)	(3.362)		(3.838)
Distance between native and acq. language	-0.372***	-0.105***	-0.038***	-0.089**	-0.042*	-0.008
	(-12.729)	(-8.178)	(-4.291)	(-2.335)	(-1.934)	(-0.104)
Literacy rate in learning countries	0.195***	0.021***	0.009	-0.143*	0.053	-0.264**
	(4.434)	(3.205)	(0.957)	(-1.898)	(1.263)	(-2.040)
Instrument (GDP ratio)		0.514***			0.328***	
		(12.118)			(4.316)	
Intercept	0.015***	0.003*	-0.006***	0.398	-0.234	2.199***
	(3.781)	(1.956)	(-2.818)	(1.385)	(-1.418)	(3.809)
No. of observations	2,502	2,502	2,502	345	345	345
(Pseudo) R-squared	0.281	0.276		0.078	0.177	
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: 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	-0.190	0.652	-0.291	0.140	0.229	0.611
Chinese	-0.244	0.473	-0.515	-0.218	0.011	-19.143
Dutch	-0.209	0.360	-0.580	0.049	0.221	0.220
English	-0.647	1.561	-0.414	0.015	0.434	0.035
French	0.005	0.795	0.007	0.035	0.208	0.168
German	-0.075	0.659	-0.114	-0.101	0.184	-0.550
Italian	-0.062	0.676	-0.092	-0.065	0.139	-0.466
Japanese[(b)]	-0.214	0.554	-0.387			
Malay	-0.070	0.253	-0.278	0.416	0.297	1.400
Portuguese[(b)]	-0.170	0.274	-0.618	-0.120		
Russian	0.079	0.597	0.131	0.050	0.252	0.199
Spanish	-0.207	1.196	-0.173	-0.021	0.226	-0.093
Turkish	-0.114	0.326	-0.352	0.045	0.186	0.239

(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	Std. 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]	0.88	0.10
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]	0.84	0.11
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