The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers

www.cesifo.org/wp

Cultural Integration and Export Variety Overlap Across Countries

Peter H. Egger
Andrea Lassmann

CESIFO WORKING PAPER NO. 4800
CATEGORY 8: TRADE POLICY
MAY 2014

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 4800*

Cultural Integration and Export Variety Overlap Across Countries

Abstract

This paper assesses the role of a larger degree of common language use between the populations of two countries on the so-called extensive product margin of trade. We focus on the overlap of products exported or imported between any pair of countries. The results suggest that the effect of varying aspects of sharing a common language on the variety overlap is both positive and important. The effect of sharing a common spoken language exceeds the one of common native language, implying that a larger overlap in language proficiency is quantitatively more important than a higher cultural proximity.

JEL-Code: C310, F140, F150, Z100.

Keywords: common language, international trade, trade costs, cultural integration.

Peter H. Egger
KOF Swiss Economic Institute
ETH Zurich, WEH D4
Weinbergstrasse 35
Switzerland – 8092 Zurich
egger@kof.ethz.ch

Andrea Lassmann
KOF Swiss Economic Institute
ETH Zurich, WEH D4
Weinbergstrasse 35
Switzerland – 8092 Zurich
lassmann@kof.ethz.ch

May 8, 2014
We thank Jan Fidrmuc, Michele Gazzola, Bengt-Arne Wickström, and participants at the CESifo Venice Summer Institute-Workshop *The Economics of Language Policy* for helpful comments. Egger acknowledges funding from GA ČR through grant no. P402/12/0982.

1 Introduction

While many models of the so-called *new trade theory* feature some elements of classical gains from trade due to endowment differences (Helpman and Krugman, 1987) and Ricardian comparative advantage associated with technology differences across countries (Eaton and Kortum, 2002; Melitz, 2003), the key paradigm of most of the models in new trade theory are gains from variety associated with products available to consumers in the open economy but not under autarky (Dixit and Stiglitz, 1977; Krugman, 1979; Feenstra, 1994; Broda and Weinstein, 2006).

Since consumers display a love of variety with (in many models) a constant elasticity of substitution between varieties as well as of demand, at the margin, when moving from autarky to the open economy, they display an infinite willingness to pay for the marginal unit of any variety that is not available. Hence, fixed costs of (bilateral) market access of exporters (Helpman et al., 2008; Egger et al., 2011) are often presumed to be the main reason of limited access to varieties of consumers across countries or, as we might call it, the missing trade in varieties.[1]

The goal of this paper is to shed light on the importance of one (fixed and variable) trade cost ingredient on the trade of varieties as measured by the extensive product margin: common language.[2] The latter is one of the most important and frequently-used arguments in the trade cost function that empirical trade economists employ when estimating models of bilateral demand in goods. Egger and Lassmann (2012) provide a large number of references and a meta analysis on the subject. Articles which focus on common language as a driver of trade are Melitz (2008), Fidrmuc and Fidrmuc (2009), Falck et al. (2012), Melitz and Toubal (2014), Sauter (2012), Egger and Lassmann (2013). To the extent that common language is a measure of common culture (Laitin, 2000; Fearon, 2003; Ginsburgh and Weber, 2013), this research is related to the role of common culture and economic exchange (Guiso et al., 2006, 2009; Felbermayr and Toubal, 2010).

This paper employs a conceptual framework for the measurement of export and import variety scope and their growth in conjunction with reduced-form analysis in order to obtain estimates of the quantitative role of common language for product variety overlap in bilateral trade. We use data on the extensive product margin of trade in the average year of $2004 - 2006$ or its change between the (averaged)

[1] In models with a variable elasticity of demand and, more generally, with non-homothetic consumer preferences, even variable trade costs may be responsible for the missing trade of varieties; see Melitz and Ottaviano (2008).

[2] The extensive product margin refers to the number of (differentiated) traded products, i.e., varieties. Note that an additional variety available to consumers, e.g., through imports, increases the extensive product margin. In new trade theory, gains from trade through the availability of new products or variety are the main source of utility gains.

periods $1994-1996$ and $2004-2006$. In particular, we analyze the overlap of products exported or imported between any pair of countries, where a product refers to any one of 5,323 Harmonized System 6-digit codes in the United Nations' Comtrade database. For the measurement of common language, we rely on a host of traditional and novel indicators as collected by Melitz and Toubal (2014): *common official language*, *common spoken language*, *common native language*, *linguistic proximity*, and an *aggregate index of common language*. Conditional on other drivers of trade (such as productivity and endowment differences, geography, and policy barriers to trade), we hypothesize that varieties exported or imported in common can be explained partly by cultural integration through using (in broad terms) a common language (due to nativeness, schooling, etc.). One issue of potential interest in this regard is a distinction between language as a mere means of exchange of information, facilitating communication in a narrow sense, and common language as one dimension of cultural proximity, which is correlated with common ethnicity and, eventually, trust (Guiso et al., 2009; Melitz and Toubal, 2014; Ginsburgh and Weber, 2013).

The results of our analysis can be summarized as follows. First, sharing a common language is important for traded product variety overlap, i.e., the set of product varieties that is traded reciprocally between two countries. We find evidence that sharing a common native language is nearly twice as important as sharing a common spoken language for variety overlap in bilateral trade across country pairs. While the former clearly reflects both cultural norms and values related to speaking a common language as well as language proficiency, we may conclude that the role of cultural proximity as measured by common native language exceeds the role of costs of translation and lack of language proficiency as captured by common spoken language. The average partial effects exceed the magnitude of other trade costs controlled for and amount to 0.345 and 0.194, respectively. Hence, an increase in the share of speakers of any given common native language by one percentage point raises the scope of common varieties traded by 0.345 percentage points for the average country pair, while a one percentage point increase in common spoken language raises it by 0.194 percentage points. Common official language seems less important (potentially due to measurement error, when interpreting it as a measure of common spoken or common native language) than geographic trade costs. The associated average partial effect amounts to 0.086. Consequently, switching from no common official language to a common official language raises trade by 8.6%. The effect of linguistic proximity is close to zero or even negative, indicating that language similarity per se is not enough to enhance trade in varieties, but what counts is having the *same* (native and, somewhat less so, spoken) language. A weighted measure of common official language, native language, and language proximity as constructed

by Melitz and Toubal (2014) and described therein as well as in Section 3 below produces results that are close to the ones for common spoken language. Taking the possibly endogenous nature of some of the measures into account leads to quantitatively lower effects regarding common native language and to higher effects regarding common spoken language and the common language index. The effect with respect to language similarity is positive and thus qualitatively different once endogeneity is accounted for. Regarding product variety *growth*, mean reversion prevails. Hence, countries with a large common language overlap as measured by either one of the indicators used in this paper start out with a *higher level* of the number of products traded but see *less of an increase* in variety overlap over time.

The remainder of the paper is organized as follows. The next section introduces measures of export and import variety as well as growth thereof. Section 3 summarizes the data used in this paper. Section 4 presents the results of the empirical analysis, including sensitivity checks. The final section concludes.

2 Measuring export and import variety scope and overlap

Let us denote countries by $i, j = 1, ..., N$, respectively, time by $t = 1, ..., T$, and individual varieties of products by $v = 1, ..., V$. Moreover, let us use $x_{t,i}(v)$ to denote the export value of variety v exported by country i, or analogously, the import variety imported by country i at time t, and let us use $x_{t,ij}(v)$ to denote the export value of variety v exported by country i to country j or the import value of variety v imported by country i from country j at time t. Finally, use $\mathfrak{V}_{t,i}$ to denote the set of varieties exported or imported by country i at time t. Then, we may follow Feenstra and Kee (2008) to define export variety as the value share of all varieties that country i and a reference country j export or import in common (to the world) as

$$\lambda_{t,i}^{j} = \frac{\sum_{v \in (\mathfrak{V}_{t,i} \cap \mathfrak{V}_{t,j})} x_{t,i}(v)}{\sum_{v \in \mathfrak{V}_{t,i}} x_{t,i}(v)}, \tag{1}$$

based on which average unilateral export or import variety scope may be defined as

$$\lambda_{t,i} = (N-1)^{-1} \sum_{j \neq i} \frac{\sum_{v \in (\mathfrak{V}_{t,i} \cap \mathfrak{V}_{t,j})} x_{t,i}(v)}{\sum_{v \in \mathfrak{V}_{t,i}} x_{t,i}(v)}, \tag{2}$$

where $(N-1)$ is the number of all countries except j and $0 < \lambda_{t,i}^{j}, \lambda_{t,i} \leq 1$.

The changes in $\lambda^j_{t,i}$ and $\lambda_{t,i}$ between two periods $t' < t$ and t may then be defined as

$$\Delta\lambda^j_{tt',i} = \frac{\lambda^j_{t,i}}{\lambda^j_{tt',i}}, \quad \Delta\lambda_{tt',i} = \frac{\lambda_{t,i}}{\lambda_{t',i}} \tag{3}$$

with $\Delta\lambda^j_{tt',i}, \Delta\lambda_{tt',i} > 0$ due to the properties of $\lambda^j_{t,i}$ and $\lambda_{t,i}$.

Analogously, we may define the value share of varieties that both i and j export to or import from each other in total bilateral exports or imports between country i and j as

$$\lambda_{t,ij} = \frac{\sum_{v\in(\mathfrak{V}_{t,i}\cap\mathfrak{V}_{t,j})} x_{t,ij}(v)}{\sum_{v\in\mathfrak{V}_{t,i}} x_{t,ij}(v)}. \tag{4}$$

Notice that $\lambda_{t,ij}$ would always be unity in the absence of one-way bilateral trade of some varieties. Whereas $\lambda^j_{t,i}$ measures variety overlap between country i and reference country j in world exports or imports and $\lambda_{t,i}$ measures average variety overlap with the average other country in world exports or imports, $\lambda_{t,ij}$ quantifies overlap of reciprocally traded varieties with bilateral one-way trade of country i with country j. As for its cousins $\lambda^j_{t,i}, \lambda_{t,i}$, it is the case that $0 < \lambda_{t,ij} \leq 1$. Moreover, the change in $\lambda_{t,ij}$ between two periods $t' < t$ and t may be defined as

$$\Delta\lambda_{tt',ij} = \frac{\lambda_{t,ij}}{\lambda_{t',ij}}. \tag{5}$$

The variables $\lambda_{t,ij}$ and $\Delta\lambda_{tt',ij}$ will be used as dependent variables in this paper. Common language variables and control variables will be used to explain these dependent variables empirically.

3 Data

3.1 Trade data

Generic levels of valued trade x are based on bilateral export and import values and quantities by 6-digit Harmonized System (HS) 1988/92 product category obtained from the World Bank's WITS database (which itself is based on the United Nations' Comtrade database). To obtain two generic years t' and $t > t'$, we utilized averaged data over the years $1994 - 1996$ for the former and $2004 - 2006$ for the latter. This strategy smoothes over cyclical movements and leads to $t \gg t'$ so that more variation is available for identification in comparison to year-to-year changes. The export variety data cover 125 exporting countries in either period, exporting to 219 countries i in both periods t' and t, and to 240 and 227 countries j in t' and t,

respectively, and 5,384 and 5,323 product categories in periods t' and t, respectively. Altogether, the raw data for the two cross-sections for t' (mid-1990s) and t (mid-2000) obtain 5,651,776 and 7,773,898 observations, respectively, on bilateral trade at the 6-digit HS level. The import variety data cover 121 countries importing 5,350 products from 240 countries in t (8,246,390 observations) and 127 countries importing 5,384 products from 227 countries in t' (6,106,289 observations). 117 countries import from 216 countries in both periods together. Clearly, not all of these raw data will be used in the analysis, since we focus on product variety *overlap* between country pairs. Hence, only those observations will be used where $\lambda_{t,ij}$ or $\Delta\lambda_{tt',ij}$ are not missing.

3.2 Common language data

In general, we use the language indicator variables published in Melitz and Toubal (2014). The data are available as a cross section and include a variety of indicators for an extensive set of countries and country-pairs. In particular, we are interested in the measures of common language: CNL_{ij}, the unadjusted sum of the products of common native language shares between countries i and j (the probability that a random pair of individuals from two countries speak the same maternal language); COL_{ij}, a binary variable for common official language between two countries;[3] CSL_{ij}, the adjusted sum of the products of the common spoken language ratios between countries i and j; $LP1_{ij}$, a normalized measure of linguistic proximity between i and j based on language trees; $LP2_{ij}$, an adjusted and normalized measure of linguistic proximity between i and j based on lexical similarity; and CLE_{ij}, an aggregate index of common language based on COL_{ij}, CNL_{ij}, and LP_{ij} together. For the construction of the measures CNL_{ij} and CSL_{ij}, Melitz and Toubal (2014) use survey information about mother tongues as well as about multiple spoken languages from the Special Eurobarometer 243 (2006). They obtain a total of 42 native and spoken languages. COL_{ij} rests upon information (extended by Melitz and Toubal, 2014) from the CIA World Factbook. $LP1_{ij}$ utilizes the Ethnologue classification of languages belonging to separate family trees, different branches of the same family tree, the same branch, and the same sub-branch. $LP2_{ij}$ is based on a similarity score from the the Automated Similarity Judgment Program between words that are common to a given language. CLE_{ij} uses the exogenous components COL_{ij}, CNL_{ij}, and LP_{ij}

[3]Some countries such as Belgium or Switzerland have more than one official language so that they have a common official language with other countries which would not have a common official language between them.

only and disregards CSL_{ij}, an aspect of language that is likely endogenous to trade.[4] For an exhaustive description of the sources and the methodology underlying the construction of the variables, see Melitz and Toubal (2014). CSL_{ij}, CNL_{ij}, CLE_{ij}, and the normalized $LP1_{ij}$ and $LP2_{ij}$ are fractional variables in the unit interval.

3.3 Control variable data

We further control for the following other determinants of bilateral trade as published by Melitz and Toubal (2014): data on the geographical distance between the two most populated cities; and binary indicator variables for contiguity, pre-World-War-II colonial relationships, and legal origin.

3.4 Descriptive statistics

The merge of the aforementioned data results in 7,084 and 6,730 bilateral pairs that have non-missing values of $\lambda_{t,ij}$ regarding export and import varieties, respectively. Descriptive statistics about the average, the standard deviation, and percentiles of data on export variety and – for those country pairs with non-missing observations of $\lambda_{t,ij}$ – data on bilateral language use are summarized in Table 1. The data clearly show that the export and language variables' individual distributions are skewed to the right. For instance, CSL_{ij} takes on a value of zero for at least a quarter of the data. It is 0.05 at the median, and it takes on a value of 0.68 in the 90th percentile of the distribution. Overall, the probability that a randomly-drawn country pair shares a common export variety, and even more so, the probability that the same pair shares a common official, spoken, or native language, or has common language roots, is rather low in general. As a consequence, there is much scope for increasing welfare gains from trade and common language overlap through acquisition of languages around the globe.

4 Empirical strategy

4.1 The effect of common language on export variety

This paper's goal is to explain product overlap $\lambda_{t,ij}$ and its change in time by standard fundamental drivers of bilateral trade as used in so-called gravity equations of bilateral trade. With panel data at hand, a host of margins of bilateral trade

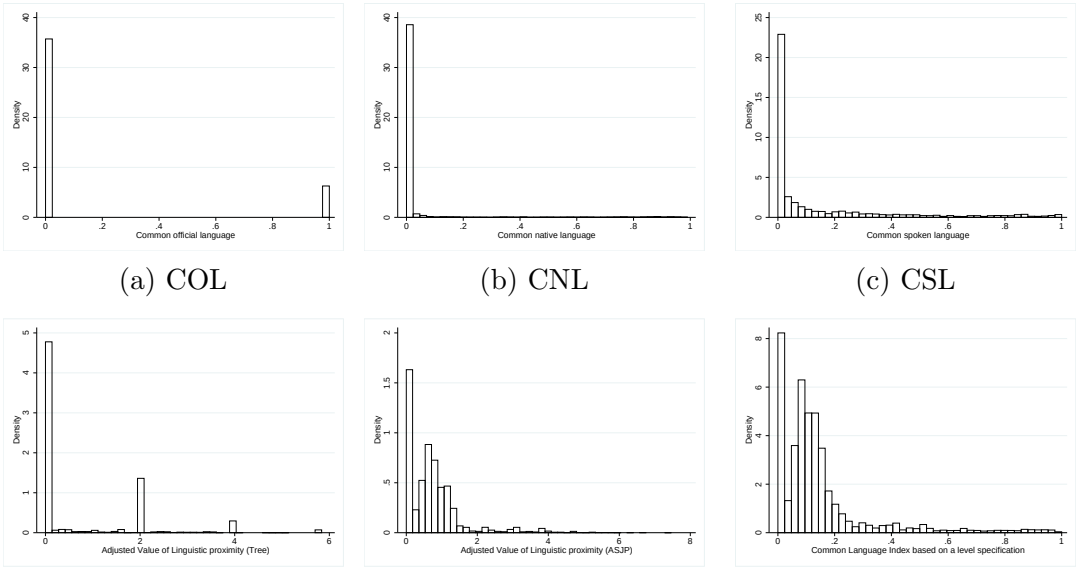
[4] This index is constructed by normalizing the sum of COL_{ij} and LP_{ij} by the maximum value and multiplying it by $1 - CNL_{ij}$; then, CNL_{ij} is added to obtain CLE_{ij}.

Table 1: Descriptive statistics

	N	Mean	Std.Dev.	p10	p25	p50	p75	p90
Exports								
$\lambda_{t,ij}$	7084	0.33	0.32	0	0.03	0.20	0.58	0.86
$\lambda_{t',ij}$	5156	0.26	0.29	0	0.02	0.13	0.44	0.76
$\Delta\lambda_{tt',ij}$	5020	57.89	852.18	0.42	1.04	1.76	5.11	26.29
Imports								
$\lambda_{t,ij}$	6730	0.29	0.32	0	0.02	0.15	0.51	0.84
$\lambda_{t',ij}$	5552	0.23	0.28	0	0.01	0.09	0.36	0.72
$\Delta\lambda_{tt',ij}$	4613	59.99	1298.30	0.58	1.09	2.06	7.02	34.50
COL_{ij}	7084	0.17	0.37	0	0	0	0	1
CNL_{ij}	7084	0.05	0.19	0	0	0	0	0.04
CSL_{ij}	7084	0.19	0.27	0	0	0.05	0.27	0.68
LP1_{ij}	7083	0.81	1.28	0	0	0	1.95	1.95
LP2_{ij}	7083	0.79	0.86	0	0.18	0.67	1.06	1.43
CLE_{ij}	7084	0.17	0.20	0	0.07	0.11	0.18	0.442

t refers to the period 2004–2006 and t' to the period 1994–1996. Variable definitions are as follows: *COL* Common official language; *CNL* Common native language; *CSL* Common spoken language; *LP1* Linguistic proximity (language trees); *LP2* Linguistic proximity (lexical similarity); *CLE* Common language index. Data sources: $\lambda_{t,ij}$, $\lambda_{t',ij}$, and $\Delta\lambda_{tt',ij}$ are based on own calculations using the United Nations' Comtrade database; all other variables are taken from Melitz and Toubal (2014).

Figure 1: Histograms of language variables



(a) COL (b) CNL (c) CSL

(d) LP1 (e) LP2 (f) CLE

– including the product margin we focus on – can be generically portrayed as a function of exporter-time-specific factors $\mu_{t,i}$ (capturing productivity, endowments, factor costs, numbers of exporters, etc.), importer-time specific factors, $m_{t,j}$ (capturing expenditures on goods and the consumer price index), and country-pair(-time)-specific factors such as trade costs, preferences, and measurement error. We model log trade costs plus log preferences as an additive function of the form $\gamma_{\text{Lang}}\text{Lang}_{ij} + \sum_{k=1}^{K} \beta_k \tau_{ij}^k$, where Lang_{ij} refers to COL_{ij}, CNL_{ij}, CSL_{ij}, LP1_{ij}, LP2_{ij}, or CLE_{ij}, and τ_{ij}^k with $k = 1, ..., K$ are observable (non-language) time-invariant bilateral trade frictions (or non-frictions) such as log bilateral distance, land contiguity, and colonial relationships after 1945. More specifically, we postulate the empirical model

$$E(\lambda_{t,ij}|\text{Lang}_{ij}, \tau_{ij}^k) = \Phi(\mu_{t,i} + m_{t,j} + \gamma_{\text{Lang}}\text{Lang}_{ij} + \sum_{k=1}^{K} \beta_k \tau_{ij}^k), \tag{6}$$

where t refers to the average of period 2004–2006, $\lambda_{t,ij}$ is the share of common bilateral export varieties. The model in (6) specifies the conditional expectation of the product overlap $\lambda_{t,ij}$ as a function $\Phi(\cdot)$ of Lang_{ij}. The purpose of that model is to estimate the parameter γ_{Lang} while controlling for observable factors τ_{ij}^k and for unobservable exporter-time and importer-time factors $\mu_{t,i} + m_{t,j}$ so that γ_{Lang} does

not reflect spurious effects from an omission of those factors.

Since $0 < \lambda_{t,ij} \leq 1$ is a fractional response variable, γ_{Lang} can be consistently estimated using a Bernoulli quasi-likelihood model (QMLE) with the standard normal cumulative distribution function indicated by $\Phi(\cdot)$. Robust inference can be obtained along the lines of Papke and Wooldridge (1996). Table 2 reports the average partial effects, i.e., the average increase in $\lambda_{t,ij}$ related to an increase by one unit (either by unity or by one percent, or by one percentage point, depending on the measure of Lang_{ij}), of an observable variable at a time.

Columns (1)–(6) of Table 2 present the results from this nonlinear empirical model including one of the language variables at a time. The magnitudes of the effects of the variables of interest on product variety overlap are as follows. The average partial effect (APE) of common official language reported in Column (1) amounts to 0.086 and, thus, sharing a common official language increases export variety by 8.6 percentage points.

As argued by Melitz and Toubal (2014) and Egger and Lassmann (2013), COL_{ij} reflects a weighted impact of CSL_{ij} and CNL_{ij} as the joint influence of both language proficiency and common cultural factors. More specifically, we expect a stronger role of the latter because it involves speaking a common language together with contextual cultural proximity aspects. Melitz and Toubal (2014) show that COL_{ij} is a rather imperfect measure of common language use, reflecting a variety of historical and political factors. Therefore, we expect it to be of lesser importance than CSL_{ij} and CNL_{ij} not only for bilateral trade in general but also for various margins of bilateral trade. Hence, we are interested in whether these two separate aspects affect $\lambda_{t,ij}$ differently or not. Columns (2) and (3) suggest that, with an average partial effect of 0.196, the effect of CNL_{ij} is more than twice as big as the one of COL_{ij}. It is also bigger than the one of CSL_{ij}, whose average partial effect amounts to 0.160 (the latter being still substantially larger than that of COL_{ij}). With 90-th and 99-th percentiles of CNL_{ij} amounting to 0.02 and 0.88, respectively, the marginal effects evaluated at these values are 0.196 and 0.221. Similarly, the 90-th and 99-th percentiles of CSL_{ij} are 0.52 and 0.97, and the corresponding marginal effects amount to 0.177 and 0.181. Hence, the average partial effects vary in the distribution of the language data due to the nonlinear functional form of $\Phi(\cdot)$ in (6). The average partial effect of language similarity as addressed in Columns (4) and (5) is close to zero, and the partial effects evaluated at different percentiles of language similarity are almost identical in magnitude. Hence, what matters for product (i.e., preference) overlap is having a large common language base but less so the similarity between two foreign languages. Finally, the APE with respect to the common language indicator as a combined measure of common language is strong and amounts to 0.214, according to Column (6). The partial effects amount

Table 2: Average partial effects of Lang_{ij} on bilateral export variety overlap, $\lambda_{t,ij}$ (QMLE)

	(1) COL	(2) CNL	(3) CSL	(4) LP1	(5) LP2	(6) CLE	(7) CNL	(8) CSL	(9) LP1	(10) LP2	(11) CLE
	(baseline specification)					(including square terms of Lang_{ij})					
Lang_{ij}	0.086	0.196	0.160	-0.004	-0.003	0.214	0.345	0.194	-0.022	-0.047	0.185
	(0.009)***	(0.017)***	(0.013)***	(0.003)	(0.004)	(0.018)***	(0.069)***	(0.031)***	(0.005)***	(0.007)***	(0.038)***
Log distance	-0.131	-0.128	-0.125	-0.141	-0.141	-0.125	-0.127	-0.124	-0.140	-0.137	-0.126
	(0.004)***	(0.004)***	(0.004)***	(0.004)***	(0.004)***	(0.004)***	(0.004)***	(0.004)***	(0.004)***	(0.004)***	(0.004)***
Contiguity	0.096	0.096	0.085	0.104	0.103	0.091	0.091	0.083	0.101	0.102	0.092
	(0.015)***	(0.015)***	(0.015)***	(0.015)***	(0.015)***	(0.015)***	(0.015)***	(0.015)***	(0.015)***	(0.015)***	(0.015)***
Colony	0.026	0.048	0.031	0.073	0.074	0.044	0.044	0.031	0.067	0.063	0.045
	(0.020)	(0.020)**	(0.019)	(0.020)***	(0.020)***	(0.020)**	(0.020)**	(0.019)	(0.020)***	(0.020)***	(0.020)**
N	7083	7083	7083	7083	7083	7083	7083	7083	7083	7083	7083

Robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Regressions include exporter and importer fixed effects (average partial effects not reported). t refers to the period 2004–2006. Variable definitions are as follows: *Log distance* Log distance between the two most populated cities in two countries; *Colony* Colonial relationship prior to 1945; *COL* Common official language; *CNL* Common native language; *CSL* Common spoken language; *LP1* Linguistic proximity (language trees); *LP2* Linguistic proximity (lexical similarity); *CLE* Common language index. The source of all explanatory variables is Melitz and Toubal (2014).

to 0.229 at the 90-th percentile ($\text{CLE}_{ij} = 0.33$) and to 0.242 at the 99-th percentile ($\text{CLE}_{ij} = 0.90$).

Overall, we conclude from the inspection of the results for the various common language measures employed in Table 2 that the effect of sharing a common language is important, it exceeds the effect of other considered trade costs, and it becomes more important the more pronounced cultural overlap in terms of native rather than just spoken (or official) language is. Finally, it is noticeable that the APEs of distance and contiguity are robust across columns (expect for those including measures of language similarity) of Table 2, whereas the APE regarding common colonial relationships is not. The latter illustrates the collinearity of cultural variables and the danger of an endogeneity when omitting relevant cultural or institutional determinants of trade. We will address issues of endogeneity regarding common language specifically in Section 4.3. We tested the robustness of the results shown in Table 2 by repeating the estimation using OLS (i.e., a linear model) and QMLE when following the Mundlak-Chamberlain-Wooldridge device by including exporter and importer averages of all variables instead of exporter and importer fixed effects (the latter might be problematic with nonlinear models, since many country-specific effects might lead to local optima of the likelihood function). The sign and the magnitude of the average partial effects presented in Table 2 turn out robust to those alternative modeling strategies.

Columns (7)–(11) allow for some degree of nonlinearity by adding quadratic terms of Lang_{ij} to (6). The corresponding APE estimates are somewhat bigger than the ones in the baseline specification in Columns (1)–(6). This is particularly the case regarding CNL_{ij}, where the APE amounts to 0.345 in Column (7) while it was 0.196 in Column (2). The APEs of LP1_{ij} and LP2_{ij} remain negative but turn out significant when considering additional nonlinearities as in Columns (7)-(11). The APE of CLE_{ij} is 0.185 as shown in Column (11), which is slightly lower than the APE in Column (6).

4.2 The effect of common language on import variety

While Table 2 focused on export flows, Table 3 undertakes the same analysis and is identically structured but uses import data. The main reason for doing so is that for some country-pairs exports may be recorded more completely (e.g., when considering exports from developed to developing economies), while for others imports may be more complete (for obvious protectionist reasons; in particular, when considering trade among developing economies). With a potentially lower average quality of reporting standards for exports, a country's product scope in trade might be underestimated, especially, for trade lines (products) where the quantities shipped are

small. This might bias the extent of language overlap. Overall, the APEs of Lang_{ij} on import variety $\lambda_{t,ij}$ in Table 3 are fairly similar to – just somewhat smaller than – the ones on export variety in Table 2. The APE is 0.079 regarding COL_{ij} (Column 1), 0.117 regarding CNL_{ij} (Column 2), 0.128 regarding CSL_{ij} (Column 3), and 0.140 regarding CLE_{ij} (Column 6). Again, the APE is close to zero regarding LP1_{ij} and LP2_{ij}, pointing to an effect on $\lambda_{t,ij}$ of those two measures that is negligible relative to the measures of common language overlap. Including square terms of Lang_{ij} increases the point estimates throughout as suggested by Columns (7)–(11). The latter is in line with the earlier findings based on exports.

4.3 Sensitivity analysis

Endogeneity bias of common language on $\lambda_{t,ij}$

The effect of common language in (6) would be inconsistent if the measures behind Lang_{ij} were correlated with the error term. E.g., this might be the case due to an omission of cross-migration as a determinant of both trade and Lang_{ij} (Melitz and Toubal, 2014). In general, COL_{ij} is very sensitive to the inclusion and exclusion of cultural, historical, and institutional variables as highlighted in Egger and Lassmann (2012), pointing to its likely endogeneity. In this subsection, we take the possibility of such endogeneity into account. We estimate the effect of common language on $\lambda_{t,ij}$ by invoking a control function approach (CF).[5]

Let us define two vectors of variables,

$$\boldsymbol{w}_{ij} = \left\{\tau_{ij}^1, ..., \tau_{ij}^K, \tau_{i.}^1, ..., \tau_{i.}^K, , \tau_{.j}^1, ..., \tau_{.j}^K\right\}, \boldsymbol{z}_{ij} = \left\{\tau_{ij}^1, ..., \tau_{ij}^{K+1}, \tau_{i.}^1, ..., \tau_{i.}^{K+1}, \tau_{.j}^1, ..., \tau_{.j}^{K+1}\right\},$$

where τ_{ij}^k for $k = 1, ..., K$ refers to any one of the trade cost variables in Section 4.1, and τ_{ij}^{K+1} refers to common legal origins as described in Section 3.3. Our two-way data are generally unbalanced. For any time-invariant generic variable v_{ij}, the generic i-specific and j-specific averages $v_{i.}$ and $v_{.j}$ are obtained from a generalized

[5]The basic idea of the control function approach is the following. Consider any right-hand side variable in an econometric model that is endogenous. Endogeneity means that part of the variable is correlated with some other variable that cannot be measured and is part of the error term. Suppose that the endogenous variable can be split into the correlated (endogenous) and the uncorrelated (exogenous) part. In practice, these two parts have to be estimated, e.g., by a regression, where the endogenous part is obtained as a residual. Then, controlling for this residual in the regression model of interest means including a control function which picks up the endogeneity bias of the parameter on the original (endogenous) measure of interest. In linear regression models, this procedure generates results that are analogous to 2SLS, however, the CF approach is able to solve the endogeneity in a number of nonlinear models such as the fractional response model applied in this paper in contrast to 2SLS; see Wooldridge (2010) for details.

Table 3: Average partial effects of Lang_{ij} on import variety $\lambda_{t,ij}$ (QMLE)

	(1) COL	(2) CNL	(3) CSL	(4) LP1	(5) LP2	(6) CLE	(7) CNL	(8) CSL	(9) LP1	(10) LP2	(11) CLE
	(baseline specification)					(including square terms of Lang_{ij})					
Lang_{ij}	0.079	0.117	0.128	-0.004	-0.004	0.140	0.212	0.208	-0.008	-0.017	0.198
	(0.009)***	(0.016)***	(0.012)***	(0.003)	(0.003)	(0.017)***	(0.066)***	(0.028)***	(0.005)	(0.007)**	(0.035)***
Log distance	-0.130	-0.131	-0.126	-0.139	-0.139	-0.129	-0.131	-0.125	-0.139	-0.138	-0.129
	(0.004)***	(0.004)***	(0.004)***	(0.003)***	(0.003)***	(0.004)***	(0.004)***	(0.004)***	(0.003)***	(0.004)***	(0.004)***
Contiguity	0.110	0.113	0.099	0.118	0.117	0.106	0.110	0.095	0.117	0.116	0.103
	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***
Colony	0.053	0.083	0.063	0.097	0.098	0.079	0.080	0.063	0.096	0.095	0.077
	(0.017)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***	(0.016)***
N	6729	6729	6729	6729	6729	6729	6729	6729	6729	6729	6729

Robust standard errors in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Regressions include exporter and importer fixed effects (average partial effects not reported). t refers to the period 2004–2006. Variable definitions are as follows: *Log distance* Log distance between the two most populated cities in two countries; *Colony* Colonial relationship prior to 1945; *COL* Common official language; *CNL* Common native language; *CSL* Common spoken language; *LP1* Linguistic proximity (language trees); *LP2* Linguistic proximity (lexical similarity); *CLE* Common language index. The source of all explanatory variables is Melitz and Toubal (2014).

projection in line with Kang (1985).[6] Let us propose the linear first-stage reduced-form model for Lang_{ij} as

$$\text{Lang}_{ij} = \pi + \boldsymbol{z}_{ij}\boldsymbol{\delta} + u_{\text{Lang}_{ij}},$$

where π is a constant, $\boldsymbol{\delta}$ is a conformable parameter vector, and $u_{\text{Lang}_{ij}}$ is a residual term. Of this model, the estimate $\hat{u}_{\text{Lang}_{ij}}$ may serve as a control function in a second-stage model of $\lambda_{t,ij} = g(\text{Lang}_{ij}, \hat{u}_{\text{Lang}_{ij}})$. The control function is based on $\hat{u}_{\text{Lang}_{ij}}$. In the second stage, we then propose the specification of $\Phi(\cdot)$ of the form

$$E(\lambda_{t,ij}|\boldsymbol{w}_{ij}, \text{Lang}, \hat{u}_{\text{Lang}_{ij}}) = \Phi(\psi + \boldsymbol{w}_{ij}\boldsymbol{\zeta} + \gamma_{\text{Lang}}\text{Lang}_{ij} + \rho_1\hat{u}_{\text{Lang}_{ij}} \times \text{Lang}_{ij} + \rho_2\hat{u}^2_{\text{Lang}_{ij}} + \hat{u}_{\text{Lang}_{ij}} \times \boldsymbol{\omega}_{ij}\boldsymbol{\theta}), \quad (7)$$

where Φ is the standard normal cumulative density function, ψ is a constant, and $\boldsymbol{\omega}_{ij} = \left\{\tau^1_{ij}, ..., \tau^K_{ij}\right\}$ is a $1 \times K$ subvector of $\boldsymbol{w}_{ij}$. Note that we allow for a high degree of flexibility regarding the inclusion of the control function in (7).[7] Inference regarding the control function yields insights about endogeneity affecting the previous section's results.

Then the average structural function (ASF) is obtained by estimating the conditional expectation in (7) and averaging out the reduced-form residuals to obtain average partial effects with respect to Lang_{ij} (see Wooldridge, 2005):

$$\widehat{ASF}(\boldsymbol{w}_{ij}, Lang_{ij}) = N^{-1}\sum_{i=1}^{N}\Phi(\psi + \boldsymbol{w}_{ij}\hat{\boldsymbol{\zeta}} + \hat{\gamma}_{\text{Lang}}\text{Lang}_{ij} + \hat{\rho}_1\hat{u}_{\text{Lang}_{ij}} \times \text{Lang}_{ij} + \hat{\rho}_2\hat{u}^2_{\text{Lang}_{ij}} + \hat{u}_{\text{Lang}_{ij}} \times \boldsymbol{\omega}_{ij}\hat{\boldsymbol{\theta}}). \quad (8)$$

In general, the measures behind Lang_{ij} are discrete or have limited support, hence a linear specification of the reduced form is only an approximation of the true underlying model and we cannot assume independence between $\boldsymbol{w}_{ij}$ and $u_{\text{Lang}_{ij}}$.

Note that COL, which is a binary variable, and CNL, which is fractional, have a large mass of data at zero. Therefore, for both COL and CNL we opt for a control function based on the standard normal cumulative distribution function in the first stage (which is based on an indicator that is unity whenever CNL is positive instead of the fractional response). For the other fractional variables, we

[6] $v_{i.}$ and $v_{.j}$ are properly centered predictions from fixed-effects regressions as outlined in Davis (2002).

[7] We performed a likelihood-ratio test of a model without quadratic and interaction terms of the control function against one including those terms. The null hypothesis was rejected. In order to avoid possible collinearity, we do not include a quadratic term of Lang_{ij} in this section.

use a control function based on the fractional prediction with multiplicative error structure similar to Wooldridge (2013) and Terza et al. (2008).[8] The residuals are replaced by the generalized residual from a probit regression in case of COL_{ij} and CNL_{ij}[9] and by the Pearson residuals in case of the fractional variables other than CNL_{ij}. The *average partial effects* based on this nonlinear approach may provide reasonable approximations compared to the ones from neglecting the discreteness or boundedness of the endogenous variables in the first stage.

In general, the coefficients are less robust than the ones in previous sections and should therefore be treated with caution. APEs of Lang_{ij} with linear reduced form in Columns (2)–(6) seem to overstate the magnitude of the effect of Lang_{ij} on export variety. Estimates based on a nonlinear reduced form in Columns (7)–(12) seem to provide better approximations. To summarize the latter, the APE of COL_{ij} turns out to be insignificant in contrast to the estimate in Table 2, amounting to -0.057. However, note that the partial effect of the control function is insignificant, pointing to either the fact that endogeneity does not pose a major challenge to our estimation regarding common official language, or to the possibility that this method is not able to solve the assumed endogeneity due to the ignorance of the specific functional form of the variables underlying Lang_{ij}. In any case, the average partial effect obtained in the previous section is preferable with regard to COL_{ij}. With a mass point at zero, functional form issues similarly arise regarding CNL_{ij}, however the control function is significant and the APE of CNL_{ij} amounts to 0.237 which is slightly lower than the previously estimated APE. All other APEs are higher in magnitude than previous estimates, and positive in case of LP1_{ij} and LP2_{ij}. An increase in CSL_{ij} and CLE_{ij} by one percent is associated with an increase in export variety by 0.704 and 0.397 percent, respectively. As the latter reflects a weighted measure of COL_{ij}, CNL_{ij}, and LP_{ij}, it is not surprising that the APE is in the interval of the ones of its constituting variables in magnitude. The former is more than twice as important as the one of CNL_{ij} in magnitude. This finding is in line with Egger and Lassmann (2013), where the relative importance of common native language on the import value and the number of transactions is about one-third of the combined effects of common native and common spoken language. The APEs of LP1_{ij} and LP2_{ij} amount to 0.183 and 0.607, respectively. The results with respect to import variety shown in Table 5 are qualitatively similar.[10] Overall, estimates that do not take endogeneity into account seem to lead to a downward bias of the true effect of

[8] We hereby replace the normalized measures of language proximity by the non-normalized measures that are in the interval between 0 and 1.

[9] $\hat{u}_{\text{Lang}_{ij}} \equiv \text{Lang}_{ij} \frac{\phi(\boldsymbol{z}_{ij}\hat{\boldsymbol{\delta}})}{\Phi(\boldsymbol{z}_{ij}\hat{\boldsymbol{\delta}})} - (1 - \text{Lang}_{ij}) \frac{\phi(-\boldsymbol{z}_{ij}\hat{\boldsymbol{\delta}})}{\Phi(-\boldsymbol{z}_{ij}\hat{\boldsymbol{\delta}})}$, where $\phi(\cdot)/\Phi(\cdot)$ denotes the inverse Mills' ratio.

[10] Note that the APE of COL_{ij} is now significantly different from zero, however, the previous objections hold.

cultural integration on traded varieties.

The effect of common language on variety growth $\Delta\lambda_{tt',ij}$

As a final robustness check we test whether sharing a common language affects export and import variety growth between two countries. While the cross-sectional nature of our language data does not allow us to shed light on the role of increasing cultural integration for the integration of the set of varieties exported or imported by a given country, such developments are rather persistent (Guiso et al., 2009), and we argue that we are thus nevertheless able to gain insights into the role of cultural integration itself on $\Delta\lambda_{tt',ij}$. We proceed with a linear regression of log variety growth on the same set of variables as used in the previous section and summarize the results in Table 6. The negative signs of the results on both export variety growth in Columns (1)–(6) and import variety growth in Columns (7)–(12) indicate that regression to the mean occurs. The coefficients on CNL_{ij}, CSL_{ij}, and CLE_{ij} are significantly different from zero for variety growth in both directions. They amount to -0.363, -0.529 and -0.527 for export variety growth, and to -0.469, -0.570 and -0.716 for import variety growth, respectively.

5 Conclusion

In this paper, we analyze the role of different measures of language for varieties traded between countries. We compare six different language measures including common official, common native and common spoken language as well as language similarity, and an index of common language. Moreover, we account for potential endogeneity of common language by applying a control function approach. Every common language measure covered captures different aspects of sharing a common language, including the ones related to cultural proximity as well as ones related to mere proficiency of speaking.

The results suggest a positive and quantitatively important effect of common language on (aggregate) product overlap in traded goods between pairs of countries. Estimates that do not account for the possible endogeneity of common language seem to understate the true effect of cultural integration on trade in varieties, and provide only an estimate of the lower bound of the impact. Both common native and common spoken language have a strong positive impact on product overlap. Once endogeneity is accounted for, common spoken language appears relatively more important than common native language. Hence, exactly that aspect which can be affected by economic, educational, and migration policy, namely common spoken language, displays relatively large economic effects. The latter opens the field for

Table 4: Average partial effects of Lang_{ij} on log export variety (control function approach)

	(1) COL	(2) CNL	(3) CSL	(4) LP1	(5) LP2	(6) CLE	(7) COL	(8) CNL	(9) CSL	(10) LP1	(11) LP2	(12) CLE
		(linear reduced form)					(nonlinear reduced form)					
$Lang_{ij}$	–	0.618	0.952	0.066	0.100	0.403	-0.057	0.237	0.704	0.183	0.607	0.397
	–	(0.100)***	(0.062)***	(0.009)***	(0.013)***	(0.052)***	(0.045)	(0.072)***	(0.063)***	(0.063)**	(0.127)***	(0.049)***
$\hat{u}_{\text{Lang}_{ij}}$	–	-0.606	-0.462	-0.051	-0.079	-0.372	-0.020	-0.139	-0.110	0.001	-0.105	-0.158
	–	(0.078)***	(0.058)***	(0.009)***	(0.014)***	(0.052)***	(0.022)	(0.030)***	(0.022)***	(0.022)	(0.039)***	(0.021)***
Log distance	–	-0.085	-0.054	-0.125	-0.126	-0.099	-0.133	-0.107	-0.068	-0.128	-0.126	-0.100
	–	(0.007)***	(0.008)***	(0.005)***	(0.005)***	(0.006)***	(0.007)***	(0.007)***	(0.008)***	(0.005)***	(0.005)***	(0.006)***
Contiguity	–	0.071	-0.003	0.087	0.079	0.056	0.091	0.075	0.018	0.088	0.081	0.058
	–	(0.028)**	(0.026)	(0.023)***	(0.023)***	(0.023)**	(0.023)***	(0.027)***	(0.026)	(0.024)***	(0.021)***	(0.024)**
Colony	–	0.016	-0.100	0.140	0.123	0.043	0.134	0.067	-0.057	0.114	0.116	0.044
	–	(0.033)	(0.035)***	(0.029)***	(0.029)***	(0.030)	(0.039)***	(0.032)**	(0.034)*	(0.030)***	(0.028)***	(0.031)
N	–	7083	7083	7083	7083	7083	7083	7083	7083	7083	7083	7083

Robust standard errors (bootstrapped with 500 replications) in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Linear reduced form and QMLE second stage in Columns (2)–(6), probit reduced form and QMLE second stage in Columns (7) and (8), QMLE reduced form and second stage in Columns (9)–(12). Variable definitions are as follows: *Log distance* Log distance between the two most populated cities in two countries; *Colony* Colonial relationship prior to 1945; *COL* Common official language; *CNL* Common native language; *CSL* Common spoken language; *LP1* Linguistic proximity (language trees); *LP2* Linguistic proximity (lexical similarity); *CLE* Common language index. The source of all explanatory variables is Melitz and Toubal (2014). All regressions include a constant, two-way Mundlak terms and interactions of first-stage residuals (not reported) with reported variables and a squared term of the residual.

Table 5: Average partial effects of Lang_{ij} on log import variety (control function approach)

	(1) COL	(2) CNL	(3) CSL	(4) LP1	(5) LP2	(6) CLE	(7) COL	(8) CNL	(9) CSL	(10) LP1	(11) LP2	(12) CLE
		(linear reduced form)					(nonlinear reduced form)					
$Lang_{ij}$	–	0.572	1.081	0.078	0.120	0.537	-0.166	0.121	0.608	0.325	0.674	0.453
	–	(0.113)***	(0.077)***	(0.008)***	(0.012)***	(0.050)***	(0.018)***	(0.067)*	(0.065)***	(0.059)***	(0.109)***	(0.053)***
$\hat{u}_{\text{Lang}_{ij}}$	–	-0.819	-0.610	-0.056	-0.074	-0.453	0.013	-0.136	-0.074	-0.021	-0.049	-0.142
	–	(0.090)***	(0.069)***	(0.008)***	(0.013)***	(0.051)***	(0.022)	(0.033)***	(0.023)***	(0.021)	(0.032)	(0.021)***
Log distance	–	-0.080	-0.045	-0.126	-0.128	-0.096	-0.145	-0.115	-0.080	-0.128	-0.129	-0.102
	–	(0.008)***	(0.008)***	(0.005)***	(0.005)***	(0.006)***	(0.007)***	(0.007)***	(0.007)***	(0.005)***	(0.005)***	(0.006)***
Contiguity	–	0.096	-0.004	0.084	0.084	0.058	0.119	0.093	0.035	0.100	0.092	0.063
	–	(0.030)***	(0.029)	(0.024)***	(0.024)***	(0.023)**	(0.025)***	(0.025)***	(0.027)	(0.023)***	(0.025)***	(0.024)***
Colony	–	0.028	-0.115	0.189	0.175	0.059	0.273	0.107	-0.012	0.159	0.157	0.073
	–	(0.039)	(0.032)***	(0.028)***	(0.030)***	(0.031)*	(0.045)***	(0.031)***	(0.034)	(0.028)***	(0.029)***	(0.032)**
N	–	6729	6729	6729	6729	6729	6729	6729	6729	6729	6729	6729

Robust standard errors (bootstrapped with 500 replications) in parentheses: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Linear reduced form and QMLE second stage in Columns (1)–(6), probit reduced form and QMLE second stage in Columns (7) and (8), QMLE reduced form and second stage in Columns (9)–(12). Variable definitions are as follows: *Log distance* Log distance between the two most populated cities in two countries; *Colony* Colonial relationship prior to 1945; *COL* Common official language; *CNL* Common native language; *CSL* Common spoken language; *LP1* Linguistic proximity (language trees); *LP2* Linguistic proximity (lexical similarity); *CLE* Common language index. The source of all explanatory variables is Melitz and Toubal (2014). All regressions include a constant, two-way Mundlak terms and interactions of first-stage residuals (not reported) with reported variables and a squared term of the residual.

Table 6: Effect of Lang_{ij} on log export and import variety growth ($\Delta\lambda_{tt'}$)

	(1) COL	(2) CNL	(3) CSL	(4) LP1	(5) LP2	(6) CLE	(7) COL	(8) CNL	(9) CSL	(10) LP1	(11) LP2	(12) CLE
	(Export variety growth)						(Import variety growth)					
$Lang_{ij}$	-0.148	-0.363	-0.529	-0.018	-0.037	-0.527	-0.219	-0.469	-0.570	-0.037	-0.049	-0.716
	(0.095)	(0.183)*	(0.148)***	(0.023)	(0.028)	(0.188)**	(0.089)*	(0.161)**	(0.124)***	(0.029)	(0.035)	(0.175)***
Log distance	0.289	0.283	0.250	0.305	0.304	0.269	0.379	0.374	0.343	0.403	0.403	0.355
	(0.046)***	(0.044)***	(0.046)***	(0.044)***	(0.044)***	(0.045)***	(0.045)***	(0.043)***	(0.044)***	(0.042)***	(0.042)***	(0.042)***
Contiguity	-0.131	-0.124	-0.087	-0.135	-0.135	-0.106	-0.182	-0.179	-0.134	-0.179	-0.183	-0.142
	(0.116)	(0.118)	(0.118)	(0.119)	(0.118)	(0.117)	(0.116)	(0.114)	(0.113)	(0.116)	(0.115)	(0.114)
Colony	-0.304	-0.340	-0.247	-0.399	-0.400	-0.314	-0.251	-0.312	-0.222	-0.401	-0.394	-0.275
	(0.105)**	(0.102)**	(0.104)*	(0.101)***	(0.100)***	(0.103)**	(0.154)	(0.141)*	(0.136)	(0.142)**	(0.141)**	(0.142)
N	5020	5020	5020	5020	5020	5020	4613	4613	4613	4613	4613	4613
R^2	0.172	0.173	0.175	0.172	0.172	0.173	0.201	0.201	0.203	0.200	0.200	0.202

Robust standard errors (clustered at the exporter level) in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS regressions include a constant and exporter and importer fixed effects (not reported). t refers to the period 2004–2006 and t' to the period 1994–1996. Variable definitions as follows: *Log distance* Log distance between two most populated cities; *Colony* Colonial relationship pre 1945; *COL* Common official language; *CNL* Common native language; *CSL* Common spoken language; *LP1* Linguistic proximity (language trees); *LP2* Linguistic proximity (lexical similarity); *CLE* Common language index. Source: Melitz and Toubal (2014).

a host of potentially interesting analyses of the relative economic benefits of multilinguality and its funding and support.

An interesting point which has to do with this paper as well as earlier work on the matter relates to the two salient roles of common language – as an aspect of trade costs and one of preferences. The data and methods here do not permit a direct distinction between the two, but future work might have data and methods at its dispose which might help differentiating between these two features. Moreover, while this paper and much of the earlier work addressed common language as a determinant of goods trade margins, other aspects of the international economy – such as services trade, foreign direct investment, or migration – might be interesting to consider.

References

Broda, C. and Weinstein, D. E. (2006). Globalization and the gains from variety. *The Quarterly Journal of Economics*, 121(2):541–585.

Davis, P. (2002). Estimating multi-way error components models with unbalanced data structures. *Journal of Econometrics*, 106(1):67–95.

Dixit, A. K. and Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *American Economic Review*, 67(3):297–308.

Eaton, J. and Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5):1741–1779.

Egger, P., Larch, M., Staub, K. E., and Winkelmann, R. (2011). The trade effects of endogenous preferential trade agreements. *American Economic Journal: Economic Policy*, 3(3):113–43.

Egger, P. H. and Lassmann, A. (2012). The language effect in international trade: A meta-analysis. *Economics Letters*, 116(2):221–224.

Egger, P. H. and Lassmann, A. (2013). The causal impact of common native language on international trade: Evidence from a spatial regression discontinuity design. C.E.P.R. Discussion Papers 9441.

Falck, O., Heblich, S., Lameli, A., and Südekum, J. (2012). Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, 72(2):225–239.

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth*, 8(2):195–222.

Feenstra, R. and Kee, H. L. (2008). Export variety and country productivity: Estimating the monopolistic competition model with endogenous productivity. *Journal of International Economics*, 74(2):500–518.

Feenstra, R. C. (1994). New product varieties and the measurement of international prices. *American Economic Review*, 84(1):157–77.

Felbermayr, G. and Toubal, F. (2010). Cultural proximity and trade. *European Economic Review*, 54(2):279–293.

Fidrmuc, J. and Fidrmuc, J. (2009). Foreign languages and trade. C.E.P.R. Discussion Papers 7228.

Ginsburgh, V. and Weber, S. (2013). Culture, languages, and economics. CEPR Discussion Papers 9357, C.E.P.R. Discussion Papers.

Guiso, L., Sapienza, P., and Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20(2):23–48.

Guiso, L., Sapienza, P., and Zingales, L. (2009). Cultural biases in economic exchange? *The Quarterly Journal of Economics*, 124(3):1095–1131.

Helpman, E. and Krugman, P. (1987). *Market Structure and Foreign Trade: Increasing Returns, Imperfect Competition, and the International Economy*, volume 1 of *MIT Press Books*. The MIT Press.

Helpman, E., Melitz, M., and Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics*, 123(2):441–487.

Kang, S. (1985). A note on the equivalence of specification tests in the two-factor multivariate variance components model. *Journal of Econometrics*, 28:193–203.

Krugman, P. R. (1979). Increasing returns, monopolistic competition, and international trade. *Journal of International Economics*, 9(4):469–479.

Laitin, D. D. (2000). What is a language community? *American Journal of Political Science*, 44(1):pp. 142–155.

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

Melitz, J. and Toubal, F. (2014). Native language, spoken language, translation and trade. *Journal of International Economics*, forthcoming.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6):1695–1725.

Melitz, M. J. and Ottaviano, G. I. P. (2008). Market size, trade, and productivity. *Review of Economic Studies*, 75(3):985–985.

Papke, L. E. and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. *Journal of Applied Econometrics*, 11(6):619–632.

Sauter, N. (2012). Talking trade: language barriers in intra-canadian commerce. *Empirical Economics*, 42(1):301–323.

Terza, J., Basu, A., and Rathouz, P. (2008). Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3):531–543.

Wooldridge, J. M. (2005). *Unobserved Heterogeneity and Estimation of Average Partial Effects*. Cambridge University Press.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge, Massachusetts: MIT Press, second edition.

Wooldridge, J. M. (2013). Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables.