STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
BE551 Degree Project in Economics
Fall 2023

The Impact of Language on Bilateral Trade

Wilma Geust (25398) and Hanna Szinai (25400)

Abstract: This paper investigates the impact of different language factors on international bilateral trade from 1996 to 2020 through gravity model analysis. Sharing a common language has been proven to facilitate trade between countries and is often included as an indicator variable in the gravity model of trade. We expand on the results obtained by Melitz and Toubal (2014), who showed that the impact of language comes from three components, by extending the time frame of their analysis, as well as implementing updated language data. We find that Melitz and Toubal (2014) results hold with an extended dataset and confirm that the main influence of language on bilateral trade comes from ease of communication. We further explore changes over time and consider how the increased access to the internet globally influences the strength of the relationship of language on trade. We are interested in understanding how increased access to the internet and the subsequent decrease in communication barriers due to globalization impact this. We find that the importance of common official and native language increases over time, but that increased internet access in exporting countries reduces the importance of common language on trade.

Keywords: Gravity, language, bilateral trade, trade models, internet
JEL: F10, F12, F40

Supervisor: Jaakko Meriläinen
Date submitted: 05.12.2023
Date examined: 18.12.2023
Discussants: Ira Kansara and Kristóf Surányi
Examiner: Johanna Wallenius

We would like to acknowledge the help of our supervisor, Assistant Professor Jaakko Meriläinen, for his invaluable assistance during the process of constructing this paper. Additionally, we extend our gratitude to Professor Farid Toubal for providing us with data we needed in order to complete our research.

Table of Contents

1. Introduction

Language, and particularly the ability to communicate, is crucial for international trade and collaboration. Increased globalization and greater volumes of international trade yearly beg the question of whether the standardization of language is becoming the norm, with English becoming the universal standard for business, trade, and education at large. Concern for the disappearance of culture and native languages has been expressed throughout the globe, prompting the UN to establish the International Decade of Indigenous Languages between 2022 and 2032 (*United Nation*s). It is safe to say that the question of the importance of languages is quite connected to the zeitgeist, and it is therefore relevant to discuss the impact of different linguistic features on society. The last two decades have seen a rapid rise in both information technology and access to that technology worldwide. The spread of smartphones, 3G and now 5G, has changed not only the personal, but also the professional and business spheres. Internationalization is at an all-time high, and global commerce is booming. People from all over the world now have tools to overcome language barriers, such as online translation websites and apps.

This paper is a replication and extension of a previous research by Melitz and Toubal (2014) called *Native Language, spoken Language, translation and trade* (Melitz and Toubal. 2014), that we build upon in terms of data and definition. The paper seminally explored the different venues through which language impacts bilateral trade, employing the famous gravity model of trade. The focus of our paper is to validate this previous research on the impact of language on bilateral trade, to see whether the role of language similarity between trading partners has changed since the turn of the century, and whether it can be attributed to the evolution of information technology. The first section of this paper will focus on reviewing the existing literature in the field, and is then followed by an exploration of the data we used. Later we discuss methodology and delve deeper into our hypotheses. Our findings can be found in the Results section, followed by the conclusion, where we also discuss potential policy implications and opportunities for further research.

In our research, we employed a panel data regression incorporating country-year fixed effects, utilizing the dataset presented by Melitz and Toubal (2014). Subsequently, we expanded our analysis by incorporating observations from years beyond the publication of their study. Our results affirm the robustness of Melitz and Toubal (2014)'s findings, demonstrating their validity on a notably larger dataset. Language remains a significant estimator of bilateral trade, with coefficients of nearly identical magnitude to the original. Our research also reveals that internet access, while modest in magnitude, is a significant factor in bilateral trade. More importantly, it significantly affects the relationship between common language factors and bilateral trade by reducing its impact, which warrants further exploration.

2. Literature Review

2.1 Background Paper

As previously stated, this paper is a replication and extension of *Native Language, Spoken Language, Translation and Trade* by Melitz and Toubal (2014). Therefore, we rely on their paper heavily, and use it as the base for our research. The aim of the original paper was to explore the aggregate impact of all linguistic factors on trade in order to obtain a more holistic understanding of the impact of language compared to the most commonly used dummy variable for common language used in the gravity model. Melitz and Toubal (2014) state that the influence of language on trade consists of three components, namely ethnic ties and trust, the ability to communicate directly, and the ability to communicate indirectly through translation and interpreters. By separating the impact into these categories, they were able to examine what aspects of language are most significant and make further inferences to policy in this area. Instead of using the generalization "official language" as the dummy in the model, Melitz and Toubal (2014) created their own classification system for common language, based on four different factors; common native language (CNL), common spoken language (CSL), common official language (COL), and linguistic proximity (LP). The LP that Melitz and Toubal (2014) used was made up of two components: LP_1 and LP_2 . They constructed LP_1 to reflect the proximity of language pairs on the language trees and branches they are located on, whereas LP_2 shows how similar the two languages are in terms of vocabulary. We will explain these variables and how we used them in later sections of our paper.

The impact of the language variables was tested first on a dataset containing 9 years of data, of which some dropped out due to missing observations. Secondly, Melitz and Toubal (2014) use Rauch's Tripartite Classification to separate bilateral trade into three categories - trade of homogenous, listed, and heterogeneous (differentiated) goods, and estimate the impact of language on these goods separately, too. They first estimate the influence of the language categories ignoring the endogenous influences [free trade agreements, common currency, cross-migrants, and any other controls that may be endogenous] on bilateral trade apart from the variation automatically included in common spoken language. They do, however, control for factors like common religion, common legal system, distance, common border, factors regarding colonization, and history of wars. We will expand upon the details in the *Description of the Data* section.

Melitz and Toubal (2014) reached the overall conclusion that the impact of all linguistic factors combined is more than double that of the typical common language dummy used in the gravity model. This conclusion is a valuable contribution to previous research on the impact of language, and

additionally allows us to understand the limitations of the gravity model, discussed in more detail in the following section. The research shows that using one dimension for language alone is insufficient for understanding its impact on bilateral trade. They infer that observing the variables in different combinations allows them to separately understand what aspects are due to communication, ethnicity, and translation. According to Melitz and Toubal (2014), the ease of communication becomes apparent when CSL is significant in the presence of CNL. Adding COL to the two would indicate that translation and interpretation play a significant role on an institutional scale. Finally, if LP is included and is statistically significant, the pair suggest that LP then reflects either the ease of obtaining translations and/or degree of ethnic rapport between groups when native languages differ. Notably, however, there is no direct evidence nor background to support this interpretation. We will outgo, in this paper, from this interpretation.

Perhaps intuitively, Melitz and Toubal (2014) found that two-thirds of the impact of language comes from ease of communication alone and has nothing to do with ethnic ties. Ethnic ties and trust come into play only when cross-migrants enter the equation, and are more relevant with regards to differentiated goods. Firstly, the impact of COL, CSL and CNL are individually examined, and they find that each variable has positive and significant results for bilateral trade. (See *Appendix 1*, Columns 1-3). More importantly, when all the variables are introduced simultaneously (*Appendix 1*, Column 5), all coefficients become smaller in magnitude, and while COL and CSL remain significant, CNL is now insignificant. Adding LP (*Appendix 1*, Columns 6-7), however, leads to the coefficient for CNL to rise and become significant again, suggesting that it is the linguistic proximity of native language that matters more. To summarize, the main results obtained by Melitz and Toubal (2014) show that linguistic influence does not hinge on one aspect; although the ethnic feature of language is important in some contexts like emigration, communication is a better general indicator when it comes to level of trade and often depends on translation and interpreters.

2.1.1. Extension paper

A more recent extension was written, which builds on the previous work of Melitz and Toubal (2014). Gurevich, Herman, Toubal and Yotov (2021) created a new extended dataset with more languages (6,534 languages in total across 242 countries) and that also contains information on the similarity of languages within countries, namely the *Domestic and International Common Language Database* (DICL) (Gurevich et al., 2021). A working paper "One Nation, One Language? Domestic Language Diversity, Trade and Welfare" was published in conjunction, and includes relevant information for our investigation. We obtain additional language data (COL, CNL and LP variables) from this new dataset to add to our investigation. The motivation behind the new research was to expand the understanding of domestic language diversity on trade, specifically focusing on the linguistic shifts that have occurred

in Canada and its welfare and economic consequences. In order to analyze this, they follow Melitz and Toubal (2014)'s method for constructing indices for the aggregate impact of language in a country. However, their common language (CL) variable is now constructed based on two indices; 1) international component (ICL) and 2) domestic component (DCL).

The authors find that controlling for shared domestic language impacts its linguistic connections to all its foreign trade partners (Gurevich et al., 2021, p.5). Through a counterfactual experiment, they find that by changing the DCL component, *ceteris paribus*, the total exports change. In summary, they exemplify how reducing domestic trade costs can lead to subsequent decreases in international trade costs (Gurevich et al., 2021, p.17). This result highlights how language ties into many of the other factors/variables in the gravity model.

2.2 The Gravity Model - a theoretical background

As our research is based on the benchmark paper by Melitz and Toubal (2014) we will also employ a gravity model, the equation of which we discuss later, in the methodology section of our paper. Therefore, it is prudent to discuss literature regarding this specific economic model. The gravity model is a popular tool for international trade analysis and is often used in empirical fields such as migration, investment, environment and more (Kabir et al., 2017). The traditional gravity model was first developed in the 1960s by a group of economists led by Tinbergen. The following model specification is the basis of the gravity model that we apply in this paper, albeit with more and different variables:

$$Trade_{ij} = \alpha \times \frac{GDP_i \times GDP_j}{Distance_{ij}}$$

From this equation, we get the linear model by taking the logarithms of the gravity equation:

$$Log(Trade_{ij}) = \alpha + \beta_1(GDP_i \times GDP_j) + \beta_2 log(Distance_{ij}) + \varepsilon_{ij}$$

The gravity model in terms of international trade between two countries shows how the volume of trade ($Log(Trade_{ij})$) is proportional to their economic mass, counteracted by their relative trade friction (in this simplified form, GDP and bilateral distance) (Baier 2020). The gravity model essentially allows science to estimate trade cost between countries, as the theory follows trade in a frictionless world. Friction in the real world manifests as trade costs, such as distance between countries, colonial ties, and similar culture. One instance where friction can arise is when languages between trading partners are different, making communication more difficult. Communication has been proven to boost trade, underscoring the importance of examining situations where effective communication is not possible or encounters obstacles. Thus, we employ the gravity model, focusing on the language aspect.

Although there is no dispute with regards to the theoretical validity to the model, its empirical application raises issues with econometric methods and statistical properties such as problems with pair-wise heterogeneity (Tzouvelekas, 2007), heteroscedasticity (Frankel et al., 1995), endogeneity, and selection bias (Baier and Bargstrand, 2007), (Mahfuz et al., 2017). These problems arise due to the cross-sectional and time nature of the data used in many of its applications. Cheng and Wall (2005) discuss how the use of fixed effects (OLS) has become a common method for unobserved heterogeneity but highlight the lack of agreement on the specification of the fixed effects (Cheng, 2005). Additionally, Kabir et al. (2017) argue that the fixed effects model is inappropriate for the application of the gravity model due to its inability to estimate time-invariant factors. Another important issue to raise with regards to the gravity model is the existence of zero trade flows. Since the model applies logarithms, it is necessary to remove zeros from the equation. This can lead to selection bias, and is an issue raised in most literature that applies to the model. Quite obviously, there is much discussion and disagreement in literature with regards to appropriate methodology when it comes to the generalized gravity model.

Gravity models have always included the language variable as a dummy variable, but there has been a lack of direct justification of how the language tied to a specific country was chosen (Melitz and Toubal, 2014, p. 20). The most common proxy today for common language in gravity model literature comes from Mayer and Zignago (2011). This measure is binary and is described as "languages spoken in the country under different definitions". The data for this dummy variable comes, similarly to Melitz and Toubal (2014), from Ethnologue and the CIA World Factbook, and is based on "the fact that two countries share a common official language" (Mayer and Zignano, 2011, p. 12). A second dummy is one based on if a certain language is spoken by at least 9% of the population of both countries.

Before returning our focus back to language, it is also pertinent to understand the existing literature with regards to the other commonly applied (time-invariant) determinants of trade in the gravity model. These include common borders, legal systems, and colonial legacies, amongst others. Linnemann's (Linnemann, 1966) paper on "An Econometric Study of International Trade Flows" is credited by Deardorff as introducing more variables to Tinbergen (1962) and Pöyhönen's (1963) initial gravity model (Deardorff, 1998). Further research has been conducted to analyze the validity of using these variables. Anderson and van Wincoop, for instance, focus on the border variable, and find that national borders reduce trade between industrialized countries by 20-50% between industrialized countries (Anderson and van Wincoop, 2003). Similarly to Melitz's work critiquing the simplified use of common language as a dummy, Anderson (1979) theorized that the empirical gravity model excludes any forms of multilateral trade resistance, which thus makes its application inaccurate (Anderson and van Wincoop, 2003). Essentially, they deal with the unexplained variation in the border variable and the famous John McCallum (1995) border puzzle. More importantly, they make a strong contribution to the

statistical issues within the gravity model, specifically developing a method that solves the omitted variable bias and comparative statistics problems that arise within the model, especially clear with the treatment of the border variable.

Moving on to another variable of interest, the colonial legacy variable has been researched by Head, Mayer and Ries. They apply bilateral trade data from 1948 to 2006 to examine how post-colonial trade is impacted by independence from the colonizer(s) (Head et al., 2011). Head et al. use the gravity model, employing dyad fixed effects, and find that independence decreases bilateral trade with the metropole as well as other countries part of that same colonial empire in the long-term (by 65% in the first 40 years of independence). They take a detailed approach, investigating especially the impacts of the timing of independence events and the level of hostility with regards to these events, and find significant results; more hostile separations led to immediate trade erosions while this pattern was not as clear otherwise. Melitz and Toubal (2014) utilize the research by Head et al. in their research when controlling for countries colonial roots.

Similarly, a less discussed variable but relevant for our discussion of languages is somatic distance. Melitz and Toubal (2018) show that somatic distance has highly significant results in the gravity model for bilateral trade even when controlling for other cultural factors such as co-ancestry, language, and religion (Melitz and Toubal, 2018). Using a European sample in 1996, they find that genetic distance is highly important in explaining trade, and factors like trust prove to be less significant when applied in conjunction. Cultural impacts on trade are central in the gravity model but are difficult to control for due to potential overlaps. Similar problems are faced with the treatment of the language variable, where as mentioned, it can be difficult to separate what aspects of the language (communication, ethnic roots, etc.) are actually the determinants of trade.

Clearly, the gravity model has been and continues to be a popular tool in empirical economic research, and a significant amount of research has been conducted with regards to its validity and theoretical background. Despite this, there is room to continue to deepen the understanding of the use of certain determinants of trade such as those discussed above.

2.3 Empirical Background

In addition to our main focus paper, numerous other researchers have investigated the impact of different linguistic factors on trade and other economic phenomena. Most stem from the gravity model and the idea of barriers to trade; essentially, the angle taken in the existing literature is largely about understanding language as either a facilitator to or obstruction of international trade. In addition to this,

increasing literature has been published on the concept of the economics of language, a theoretical framework developed in 1965 by Jacob Marshak which studies language as a tool in human economic activities as well as the economic characteristics of language (Marshak, 1965, as cited in Zhang and Grenier, 2012). Research into the economic role of language became increasingly popular as a result of the official language question in Canada, which brought forth the notable differences in incomes between Anglophones and Saxophones, resulting in the birth of new strands of data collection (Zhang and Grenier, 2012.). Literature on the relationship between language and socio-economic status, the development of language policy, and the relation of language with human capital theory, amongst other concepts, became relevant amongst scholars. However, research on the topic remains fragmented, with strands of research making it difficult to understand what the field actually focuses on and what interconnections can be made (Zhang and Grenier, 2012.).

Seeing as the existing work on the impact of language on economic phenomena is widely dispersed, we focus solely on the literature surrounding bilateral trade, and disregard the impact of language on other aspects of the economy. We found that Havrylyshyn and Pritchett (1991) were among the first to apply a clear language dummy in their investigation of "European Trade Patterns After the Transition" (Havrylyshyn and Pritchett, 1991). In the paper, they research the expected change in geographical direction of exports and imports as a result of the separation of East and Central European states from the Soviet Union. They decide to use language as their proxy for cultural similarity, where the variable equals 1 if countries share a language. There is no more information on how this was determined, other than that they include separate variables for English, Spanish, Portuguese, and Arabic. Quite intuitively, they found that (with the exception of Arabic), "sharing a common language raises bilateral trade substantially" (Havrylyshyn et al., 1991, p. 6). Similarly, Foroutan and Pritchett (1992) apply the gravity equation to make inferences about trade in Sub-Saharan Africa and utilize a common official language dummy (Foroutan, 1992). Frankel, Stein and Wei (1993) continue with this pattern in their research of continental trading blocks; Europe, the Americas, and Pacific Asia (Frankel et al., 1993). The paper focuses on the impact of free-trade agreements on intra-regional trade, and includes common language and "other historical tie" dummies to control for regional biases. Once again, they find that sharing these features significantly increases bilateral trade (in this case, by 65%).

It is clear that the early works that applied language in the research of bilateral trade patterns were rather simplified and often remained unexplained. Language tended to also be used to represent culture overall in the gravity model, which even Havrylyshyn and Pritchett admit to be crude (Havrylyshyn et al., 1991). The first to define a stricter use of language in the gravity model was Rose (2000), who based the dummy strictly on the official status of the language. This definition had been lacking from the previous literature to a large extent. Rose (2000) paper on the impact of common currencies on trade includes many of the same controls that Melitz and Toubal (2014) included, but simplify language to

plainly that of official status (Rose et al., 2000). Inspired by the simplified use of language in the gravity model, Melitz was the first to crack down on the channels through which language impacts bilateral trade (Melitz, 2007). He was the first to construct separate series for language that distinguish between ease of communication, institutional status, and the role of translation, which are all different aspects of language similarity. Melitz's and Toubal's work can thus be considered seminal in the area of language and trade research. Egger and Lassman also expanded on the use of language in their meta-analysis on the language effect in international trade (Egger et al., 2012). Analyzing 701 language coefficients obtained from a range of articles published between 1970 and 2011, they find that the meta-regression reveals a direct 44% increase in trade-flows due to common (official or spoken) language. Importantly, they find that the coefficient becomes higher over time, indicating that with more recent data, the language effect on bilateral trade is larger. We investigate this in the section *Extension results and analysis*.

More recent work includes a paper published by Jan Fidrmuc and Jarko Fidrmuc in Empirical Economics, where the effect of knowledge of foreign languages on trade was analyzed by combining gravity models with data on fluency in the main EU languages between 2001 and 2007 (Fidrmuc and Fidrmuc, 2016). Once more, the gravity model is utilized and it was concluded that greater density of linguistic skills translated to greater trade intensity, where the causality between trade and language proficiency could go either way. This paper built on previous research by also including secondary speakers in their data. Unlike Melitz and Toubal (2014), the impact of language on trade was focused solely on the communicative probabilities between two set countries (the probability of two randomly chosen individuals from two countries being able to communicate), so the focus is once again on the communication aspect of language.

Another paper called *The influence of language similarity in international trade: evidence from Portuguese exports in 2013* (Ribeiro and Ferro, 2016) analyzes communication costs for trade caused by language barriers. They studied the relationship between the volume of exports from Portugal to its main 56 trading partners based on language family/language used. This paper is particularly relevant to our study, as it largely applies a similar approach to us (explained further under *Methodology*), with the use of multiple linear regression, the gravity model and many of the same variables as we do. They include real GDP, distance, and dummy variables like common language, common border, belonging to a free trade agreement, and the existence of colonial relationships. The regression shares similarities to the previously discussed papers, however, the scale of the research is only Portuguese exports during a very limited time period, the duration of a year, which risks an inaccurate and non-holistic result for the overall impact of language on bilateral trade. Moreover, their classification of language families was very limited, divided into only three categories, which leaves a notable gap as 142 language families

exist in total. Interestingly, they also decide to include free-trade agreements as a control in their regression, which Melitz and Toubal (2014) omit due to endogeneity.

Despite these limitations, the paper is valuable for our contribution. They conclude that the effect of having the same language in two countries is similar to that of sharing a culture or legal system, and that language can impact the choice of where to export. However, no connection was found to the volume of trade once the decision to export to a specific country was made, which is a point of consideration in our research. They also found that sharing the language family/language similarity (chosen based on the official language of the countries) positively and significantly impacts exports to these countries. Our research will further expand on this conclusion, exploring a dataset with a bigger scope to understand whether these results hold on a global scale and during a much longer timeframe.

Although no seminal papers have been published in the past few years with regards to language in the gravity model (other than Toubal's extension), there has been increased discussion on the role of translation. Specifically, Melitz and Toubal (2014) released an article on "The potential impact of machine translation on foreign trade" (Toubal and Melitz, 2019). The article discusses the expectation of reduced trade barriers due to both physical and linguistic distance, but does not make any definitive conclusions. Responding to Richard Baldwin (2018), according to Melitz' and Toubal's previous research, "the net result of reducing linguistic frictions with a set of trading partners is not apparent" (Toubal and Melitz, 2019). It is important to also mention one of the few works done on the exact topic: Brynjolfsson et al. (2018) showed that the introduction of machine translation to eBay in 2014 increased US sales to Latin America by 17.5% (Brynjolfsson et al., 2018). Melitz and Toubal (2019), however, suggest an air of caution with regards to this statistic as there is high risk of missing controls or information. On the other hand, Kitenge and Lahiri (2021) investigated the interaction between internet and language similarities in international trade and found that language elasticity on trade is smaller with increased internet access, a relationship that would suggest otherwise to Toubal & Melitz (2019). This discourse reveals a gap in literature and opens up for new areas of research in the field of language and bilateral trade, which we will begin to explore in this paper through further investigating the impact of internet access. Overall, there is still space to expand our knowledge on the impact of language by exploring more controls and better econometric specifications.

3. Description of the Data

In order to match our benchmark paper as closely as possible, with the goal to verify the original findings, we aimed to use the same datasets that were used by Melitz and Toubal (2014), but augmented

with data from the years that have been added since. Therefore, we will now begin with explaining the data that the original authors used.

Melitz' and Toubal's study is conducted on the basis of ten years of trade data (1998 to 2007) from 224 countries, while the construction of the four language series varied according to availability of data. For *Common official language* (COL), the CIA World Factbook was used to identify the official language(s) of the countries in the data set, with a total of 19 languages used. It is important to note that there is a limitation to a maximum of two official languages per country, which can result in inaccuracies. Secondly, for *Common native language* (CNL) and *Common spoken language* (CSL), the data was largely based on a survey conducted in the EU between 2001 and 2008 which included questions about what languages citizens spoke and at what level of confidence. Finally, *Language Proximity* (LP), was largely constructed using *Ethnologue* and uses each country's native language even if said language has no relevance outside the country. LP was then divided into two separate measures, LP_1 and LP_2. LP_1 is based on the linguistic proximity of language trees, including their classification between trees, branches, and sub-branches. The languages were then given a score between 0 and 1 based on how close or far they were on the language trees. This variable provides us with a concise and uniform method for implementing language roots into our research.

However, it does have some limitations, namely when comparing two languages that belong to completely separate trees, as this would automatically call for a linguistic proximity score of 0. Additionally, the score assumes that the LP score 0.5 means the same between different family groups, for example between the Indo-European group and the Altaic, Turkic one. Melitz and Toubal (2014) introduced the second variable LP_2 to correct for the above mentioned problems - instead of linguistic family, it relies on the Automated Similarity Adjustment Program (ASJP) score of similarity between 40 relevantly identified words. So, instead of distance between families, the focus of LP2 is on the proximity of linguistic features, which is more easily connected to ease of communication and the COL variable. Since our research is meant to validate the findings of Melitz and Toubal (2014), we will use both of these variables. However, it is worth mentioning that in his extension paper, Toubal later only used LP_1 (together with CNL and COL) to construct his new, holistic language variable CL (common language). We debated using this variable instead of Melitz' language factors, but ultimately decided to employ the old version, in order to replicate the original research. This allows us to attribute all potential differences to the increased trade data.

3.1. The Trade Data

The bilateral trade data we use is the same as that used in Melitz and Toubal (2014)'s paper, but with an extended timeframe. The data was obtained from the CEPII database and consists of the BACI

dataset (Conte, Cotterlaz and Mayer. 2022). This dataset was built (and is being updated from) the United Nations Statistical Division (commonly known as the Comtrade dataset). It is based on reported imports and exports for each dyadic country pair. These two values for each trade flow - the export of country i and the import of country j and vice versa - should be identical, but in many cases they are not, due to inconsistent reporting and data collection issues. The BACI dataset corrects this by calculating the true ratios between the import and the export value for each value, then regressing it on gravity variables and a median value that is specific for each product. The availability of data has expanded and we are now able to implement data from 1995 to 2021, as opposed to the previous research that only covered the time period 1998-2007, which was the only data available at the time. This allows us to understand whether the results hold with seventeen years of new data, and hence identify factors that may have impacted the findings.

During the data processing and analysis, the trade data we used posed some problems. There is a large number of missing values that prevented us from running the regressions we deemed necessary for our analysis. Therefore, we had to remove all observations (country-pair/year, one row) that had missing values. This is not only problematic because the dataset now ranged only from 1996 to 2021, but more importantly because it seems like there are no zeros in the dataset. Logically this cannot be, since it is reasonable to assume that a country with protectionist policies - Sudan, for example - does not trade with every country. This means that the missing zeroes might cause our data to be skewed. Melitz and Toubal (2014) also ran into this problem, but they showed that it was dismissible, by also running their regression on a sample of 50 countries with the highest GNP, assuming that it is unlikely that these countries do not trade with each other at all. This means that any NAs in this sample are likely real NAs and not zeroes. Their results for this regression indicates that it is safe to assume that the missing zeroes do not significantly affect the model and the regression.

There is another difference between the data used in the benchmark paper and ours, that needs a more detailed explanation. The BACI dataset is differentiated by type of goods traded (based on Rauch classification): homogenous, listed and heterogeneous (differentiated) goods. Due to limited resources, this paper does not differentiate between types of goods traded. In contrast, in the original paper, Melitz and Toubal (2014) did. (Due to the original authors dropping some observations, the two datasets are slightly different). They also dropped all observations that did not fit into the Rauch classification. They did this because of the assumption that trade of differentiated goods would require a higher level of communication. This was proven to be right by their investigation, where they ran the specified regression on the different types of goods. The results showed that for the trade of differentiated goods, COL is highly significant, meaning that translation plays a significant role in that particular aspect of trade. The benefits of conducting analyses based on these different product groups is thus varied and even has potential policy implications: Understanding the different needs for language and access to

translators would allow developing countries to plan according to their basket of traded products (i.e., the types of goods they trade).

3.2. The Controls

The majority of the data obtained for our control variables was collected from the CEPII Gravity Database, which encompassed many of the variables Melitz and Toubal (2014) used. Perhaps the most straightforward variable to obtain was *Contiguity*, which is simply a dummy that equals 1 if countries share a common border. Our *Distance* variable also comes from the CEPII database, and it is derived from the geodesic distance between the most populated cities in kilometers. Next, to reflect different aspects of colonialism, we used two variables. Melitz and Toubal (2014) obtained their data from Head et al. (2010) for both colonial variables (twentieth century and earlier) but we decided to use a variable that is included in the CEPII database, due to data processing reasons. It reflects the *Ex colonizer/colony* relationship with a dummy variable which is 1 if a country pair was ever in a colonial or dependency relationship (including before 1948). Similarly, for *Common colonizer*, we use a dummy that is 1 if the pair ever had the same colonizer (including before 1948). We acknowledge that the use of another database than the original authors did may slightly alter our results but should not have a significant impact on the general direction, magnitude and significance of our coefficients.

For *Common legal system*, Melitz and Toubal (2014) obtained their data from JuriGlobe and created their own dummy. On the other hand, we used the already existing legal variables in the CEPII database for data processing reasons. The CEPII database contains two different variables for this factor, namely "comleg_pretrans" and "comleg_posttrans". The two variables look at the legal system of the country before and after 1991, as the two variables might reflect different aspects of preference for trade. The cutoff point is 1991, as the fall of the USSR set off a chain reaction of countries changing their constitution and legal systems. If the country had a common legal system before 1991 it probably has no direct effect on trade between 1996 and 2020. However, it might signify a level of historical partnership that might be reflected in modern day trade levels. If the countries in question share a legal system currently (as in have the same legal system post 1991), it is reasonable to assume an ease of transaction that would elevate trade between the two countries. Therefore, it is relevant to look at both variables; however, in order to replicate the original study, we used the variable that most accurately represents the current legal situation in each country, which is the Common legal system after 1991.

A potential issue comes from our *Common religion variable*. Melitz and Toubal (2014) constructed their own variable for common religion, obtaining data from the CIA World Factbook, the International Religious Freedom Report (2007), the World Christian Database (2005) and the Pew Forum (2009). By doing this, they were able to construct a more detailed variable for common religion that takes into

account the sum of population shares that share the same religion within countries. Unfortunately, we did not have access to their exact data, and thus had to rely on the "comrelig" or religious proximity index variable from the Gravity database.

There is another variable that requires more explanation, namely *Years at War*. It reflects whether a country pair had been at war since 1823 as a measure of ancestral links and hence stands as a proxy for trust between countries. Similarly to Melitz and Toubal (2014), we were unable to find other suitable variables to reflect the history of trust and ethnic ties, and thus resorted to using the same data as they did. We obtained the data from Correlates of War dataset, using the Interstate War dataset from 1823 to 2003. Due to changes in regions and country names over time, we identified past states and areas with their corresponding modern names. In addition to identifying all former German states as Germany and former Italian kingdoms as Italy (as Melitz and Toubal (2014) did), we also replaced the USSR with Russia and the Ottoman Empire with Turkey. In order to add the variable to our aggregate data, we manually transformed the dataset from its original format into one that is aligned with the rest of the data, in order to merge the two. During the process some observations had to be dropped, meaning that the resulting data is subject to human error. We also need to disclose that Melitz and Toubal (2014) wrote in their paper that the number of years at war they used ranged from 0 to 17, whereas ours range from 0 to 15. This might be due to different processing of the data, but it is difficult to find the source of the issue, since the original paper does not go into detail about the data for this variable. However, we are confident that this difference does not influence the validity of our replication in a significant way, as both the inclusion and exclusion of our Years at War variable yields a similar result for our regressions (see section *Results and Analysis*).

Later in our research we also add a variable to analyze the effect of access to the internet on bilateral trade and language. To do this we obtained data from the World Bank database, showing how the number of people using the internet has evolved over time per country, as a percentage of the population (World Bank, 2023). The source of the data is the International Telecommunication Union, and it mainly comes from operators, household data, and business surveys. A potential issue with this data, as is with all others, is the unevenness of reporting between countries.

4. Methodology

4.1 Hypothesis

The purpose of our paper is dual; we want to verify that the original findings of Melitz and Toubal (2014) hold when subjecting their research to a much larger sample, and we also want to look more

closely at the importance of common language as a predictor of bilateral trade, and how it might have changed since the publication of our benchmark paper. Our first hypothesis therefore is that the results found by Melitz and Toubal (2014) are corroborated and verified on an expanded dataset.

When looking at the role common language plays into bilateral trade, we also looked at how using internet access influences and is influenced by the language variables in order to corroborate our theory that actively speaking a language has become less important with the emergence and diffusion of smartphones, and the widespread access to the Internet. As Melitz and Toubal (2014) pointed out, two-thirds of the influence of language upon bilateral trade comes from ease of communication (Melitz, 2012). However, the rise of free online resources that can instantly translate content from one language to the other, may decrease the importance of direct communication, and the direct effect of speaking the same language. International communication is more often than not conducted through online messaging platforms such as e-mails, which allows both parties to get help from the internet in terms of translation, which speeds up the translation processes. Growing internet access decreases the power of common language to drive trade between countries, and its ability to ease communication between economic actors (Kitenge, 2021) There is also evidence showing that the rise in access to 3G internet caused changes in interpersonal communication, global banking, information access, and more (Manacorda et al., 2020). Our theory is that internet access will lower the significance of all language variables, so that the choice of trade partner now truly only reflects historical and ethnic preference for trade, rather than ease of communication. Based on this, the second part of our hypothesis is that the evolution of information technology and the spread of the internet will impact the relationship between language and bilateral trade in a negative way.

4.2 Our Model

Firstly, we need to mention that we do not differentiate between types of goods traded; trading more complex differentiated products requires more communication and explanation (and thus benefits more from common language) than for example trading only primary goods. Several papers in the literature differentiate based on goods, but we simply do not have the time and resources to do so. To stay as close to the benchmark paper as possible, we aim to use the same variables Melitz and Toubal (2014) used, but applied on a different, and more importantly, much larger dataset. There are some variables that are often used in the literature, but that the original authors chose to exclude. In order to stay true to their work, we also did not employ these variables, however, the original authors' reasoning for excluding them is worth mentioning. Melitz and Toubal (2014) made the assumption that all control variables are exogenous, but they also examined two endogenous variables that are widely used in existing literature on the topic; free trade agreements and common currency areas. Coupled with the fact that these two variables are endogenous, and that they have no effect on the language variable if

included or excluded, the authors dropped the variables from their final regression. Melitz and Toubal (2014) also recognized the need to separately investigate the endogenous effect of cross-migration, as it is clearly related to language. The separation served a purpose of allowing the authors to reach an estimate of linguistic effects where the only endogenous variable was common spoken language. As the investigation into the effect of cross-migration was not part of the main research, and is more related to ethnic ties and trust rather than ease of communication, we decided to forgo the replication of that part of their paper, and eliminate that variable as well.

Based on the previous research, we thus formulated the following regression to analyze the significance of language on bilateral trade. We are employing a panel data regression with country-year fixed effects. We specify our model in the following manner:

$$log\ T_{i,j} = \beta_0 + \delta_c + \beta_1 COL + \beta_2 CSL + \beta_3 CNL + \beta_4 LP_{1,2} + \beta_5\ log\ D_{ij} + \beta_6 Cont_{ij}$$
$$+ \beta_7 ExCol_{ij} + \beta_8 ComCol_{ij} + \beta_9 ComRel_{ij} + \beta_{10} ComLeg_{ij} + \beta_{11} YearsAtWar_{ij} + \varepsilon_{ij}$$

Where the dependent variable T_{ij} is the level of trade between countries i and j, COL, CSL, and CNL are the language indices developed by Melitz and Toubal (2014) in their paper, LP is the subindex variable for language proximity as defined by Melitz and Toubal (2014). LP_1 is derived from *Ethnologues* language trees, and obtained from the United States International Trade Commission's *Domestic and International Common Language Database* (DICL), while LP_2 is based on the ASJP scores of languages. δ_{ij} stands for the country-year fixed effects we use to encompass all unobserved variation that is country or time specific. The control variables used are as follows; D_{ij} represents the distance between country i and j, $Cont_{ij}$ shows whether the country i and country j are adjacent or not, $ExCol_{ij}$ represents ex colonies or colonizers, $ComCol_{ij}$ stands for common colonizer, $ComLeg_{ij}$ represents the existence of a common legislative system, and $ComRel_{ij}$ is the existence of a common religion. The justification of the use of these variables are worth looking into in more detail.

Distance has recently posed an interesting paradox to the scientific community. The variable has long been used in literature and research on bilateral trade as a proxy for transportation costs and many would assume that in our highly globalized world the role of distance would diminish over time, but it has not shown to be so. Most scientists who examine this variable have come to the conclusion that distance is still relevant, the explanation of which Coe et al. (2002) tried to specify in their paper "The Missing Globalization Puzzle" (Coe et al., 2002). They found that the distance variable can be biased due to the missing zeroes in the data, the missing concept of "multilateral trade resistance" in the gravity model of trade, or the misspecification of the model stemming from omitted variable bias (distance might not

be a sufficient estimator for transport cost). Due to these possible reasons, distance remains significant in most models, even though it is counterintuitive.

Contiguity, Common legal system, Common religion and *Colonial ties* are commonly used in gravity models to account for trade barriers. *Colonial ties* especially have been shown to significantly lower the unexplained variation of gravity model regressions in the literature (Egger, 2012). The *Years at War* variable is meant as a stand-in for dyadic trust and cultural links between the countries. Ideally, Melitz and Toubal (2014) would have used a trust survey, such as that conducted by Guiso et al. in 2009 in their paper "*Cultural Biases in Economic Exchange*". This paper was limited to culture and trade in Europe, and explored how economic factors like investment were impacted by bilateral trust levels. In their study, trust included characteristics such as cultural aspects, history of conflicts, as well as religious, somatic, and genetic similarities. Due to the limited scope of the study, Melitz and Toubal (2014) chose to use only the history of conflict aspect as an indicator of trust, and further limited this data to conflicts after 1828 as opposed to wars since 1500.

5. Results and Analysis

For a holistic analysis, we ran several different regressions and analyzed the validity of our results.

5.1 Replication Results and Analysis

In order to compare our extended data with the original results, it was essential to first replicate Melitz and Toubal (2014)'s primary findings. This is done to ensure that our model is well specified and that our dataset is sufficiently similar to that of Melitz and Toubal (2014). If the results are largely the same, with minor differences in the magnitude of our coefficients due to, in some cases, use of different datasets, we can rely on them to be a trustworthy base for further expansion.

Melitz and Toubal (2014) ran several regressions (see *Appendix 1*) where they included the different language variables individually to gauge how they act and interact with each other. They did this because one of their main aims was to examine how important different aspects of language are in terms of trade. Our paper does not dive as deep into that topic, therefore our replication regression includes all language variables, which is displayed in Table 1 below. (See *Appendix 2* for our replication regressions with the language variables on their own).

Table 1: Common Language and Bilateral Trade: Replication Regression

	Dependent variable:		
	Bilateral trade (log)		
	(1)	(2)	(3)
Common official language	0.383***	0.396***	0.386***
	(0.023)	(0.023)	(0.023)
Common spoken language	0.463***	0.419***	0.451***
	(0.039)	(0.040)	(0.040)
Common native language	0.286***	0.335***	0.308***
	(0.054)	(0.057)	(0.057)
Linguistic proximity (1)	0.132***		0.122***
	(0.006)		(0.009)
Linguistic proximity (2)		0.158***	0.019
		(0.009)	(0.014)
Distance (log)	-1.473***	-1.473***	-1.472***
	(0.007)	(0.007)	(0.007)
Common border	0.700***	0.716***	0.700***
	(0.034)	(0.034)	(0.034)
Ex colonizer/colony	1.026***	0.992***	1.024***
	(0.043)	(0.043)	(0.043)
Common colonizer	0.678***	0.663***	0.677***
	(0.018)	(0.018)	(0.018)
Common religion	0.305***	0.347***	0.305***
	(0.025)	(0.025)	(0.025)
Common legal system	-0.100***	-0.107***	-0.101***
	(0.012)	(0.012)	(0.012)
Years at war	-0.071***	-0.071***	-0.071***
	(0.019)	(0.020)	(0.019)
Observations	202,709	202,709	202,709
R^2	0.734	0.734	0.734
Adjusted R^2	0.734	0.734	0.734
Residual Std. Error	2.049 (df = 202338)	2.050 (df = 202338)	2.049 (df = 202337)
F Statistic	1,512.438*** (df = 370; 202338)	1,510.766*** (df = 370; 202338)	1,508.373*** (df = 371; 202337)

Note: Data are from Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014) and Sarkees, Meredith Reid and Frank Wayman. (2010).
Standard errors are reported in the parentheses. *p **p ***p<0.01

The reason for having three columns, representing three different regressions in Table 1, is that there are two variables representing Language proximity. One regression only contains LP_1 (*Table 1, column 1*) one only LP_2 (*Table 1, column 2*) and one with both LP_1 and LP_2 (*Table 1, column 3*). Melitz and Toubal (2014) also did this and noted; LP_1 and LP_2 were extremely similar in magnitude and significance, and the precision of the two variables was varied. They decided to use LP_2 in further regressions, since it was more intuitive to them. However, we wanted to examine both variables, which led to us running multiple regressions. It is important to note that when we included both LPs, LP_2 was

not significant in our regression. It became significant only when we removed LP_1, indicating that LP_2 on its own is a suitable measure of linguistic similarity, but that its effect is diminished by the importance of linguistic roots. We suspect that this is because many linguistic features captured in LP_2 come from the language's early roots and are thus covered in the variation of LP_1.

We find that all three replication regressions yield similar results to those in the original paper, which we will analyze further in the following section *The Language Variables*. Nearly all coefficients are of similar magnitude, have the same sign, and are of the same significance level as the benchmark, with the exception of one: common legal system. We will discuss this in the *Control variables* section.

5.1.1 The Language Variables

In Melitz and Toubal (2014)'s regression (see *Appendix 1, column 6*) with only LP_1, the coefficients for Common official language was 0.360, Common spoken language was 0.399, and Common native language was 0.294, whereas the same variables in our regression (see *column 1, Table 1, found above*) had the coefficients 0.383, 0.463, and 0.286, respectively. LP_1 was originally 0.073, whereas in our regression the coefficient is 0.132. All variables are significant at a 1% significance level.

In the original regression with only LP_2 (see *Appendix 1, column 7*), Common official language was 0.351, Common spoken language was 0.396, and Common native language was 0.284, whereas in our replication the same variables had the coefficients 0.396, 0.419, and 0.335, respectively (see *column 2, Table 1, found above*). LP_2 was originally 0.078, where in our regression the coefficient is 0.158. All variables are significant at a 1% significance level. Overall, we see that our regression produces slightly higher coefficients for all language variables, except for Common native language in the case of using LP_1 instead of LP_2.

To summarize, the language variables show a strong similarity to the original results found by Melitz and Toubal (2014), with similar signs, magnitude and significance. This gives us confidence to continue on with our analysis.

5.1.2.The Control Variables

In the original regression (*Appendix 1, column 6*) with only LP_1 present, Distance was -1.364, Contiguity was 0.662, ExColony was 1.500, Common colonizer was 0.775, Common religion was 0.264, Common legal system was 0.209 and Years at war was -0.382. In contrast, our results for the same regression (*column 1, Table 1*) were -1.473, 0.700, 1.026, 0.678, 0.305, -0.100 and -0.071, respectively. All variables are significant at a 1% significance level.

Looking at the regressions with only LP_2, we see that the original values (*Appendix 1, column 6)* for Distance was -1.365, Contiguity was 0.670, ExColony was 1.484, Common colonizer was 0.779, Common religion was 0.289, Common legal system was 0.217 and Years at war was -0.382. In contrast, our results for the same regression (see *column 2, Table 1)* were -1.473, 0.716, 0.992, 0.663, 0.347, -0.107 and -0.071, respectively. All variables are significant at a 1% significance level.

The adjusted R-squared for both regressions is as high as 73.4%, which is very close to the original 75.7%, and the F statistic is significant at the 1% significance level. The minor differences between our version and the original can be explained by the slightly different data we had to work with. Overall, we can conclude that our replication of Melitz and Toubal (2014)'s paper is similar enough so that we can enter the next stage of our analysis, confident that any further results are based on the right dataset and model.

We see a bit larger difference between the original results and ours, in the case of Common legal system and Years at War. The latter being significantly smaller than the original might be attributed to a difference in how we processed the data for that variable. Since the sign and significance remain the same we proceed, but with caution.

Due to our Common legal system behaving in an unexpected way, we specified another model that included both legal variables discussed earlier. Our results (*see Table 1*) show that if countries currently have a common legal system, it actually influences trade negatively, with a coefficient of -0.10 in both the regressions with LP_1 and LP_2 separately. This seems unintuitive to us, and therefore we investigated this further by including a variable that shows whether the countries had a common legal system before 1991, to see whether that too would be negative. These results are shown in Table 2 below, where both legal variables and both LP variables are included. Here we can see that having a common legal system in the past (before 1991) is actually positive and significant, with a coefficient of 0.517, whereas having a common legal system after 1991 is negative and significant with a coefficient of -0.471. This would suggest that a current common legal system is a deterrent of trade, in that it fails to ease business transactions. In contrast, if the countries shared a common legal system in the past, it influences trade positively, implying that historical ties are more important for choice of trade partner than ease of transaction. It is also possible that ease of transaction is reflected in another variable instead. This is a puzzling development, and seeing how Common legal system was positively significant in Melitz and Toubal (2014)'s regression, we refrain from drawing definitive conclusions regarding this variable. It is, however, worthy of note that if we interpret the interaction between the legal variables in the way described above, it would coincide with how our two language proximity variables behave. Both of

these factors seem to indicate that historical and cultural ties are more important than current relationships, in contrast to what Melitz and Toubal (2014) found in their results.

Table 2: Common Language and Bilateral Trade: Replication Regression with all variables

	Dependent variable:
	Bilateral trade (log)
Common official language	0.340***
	(0.023)
Common spoken language	0.433***
	(0.040)
Common native language	0.312***
	(0.057)
Linguistic proximity (1)	0.127***
	(0.009)
Linguistic proximity (2)	0.001
	(0.014)
Distance (log)	-1.468***
	(0.007)
Common border	0.657***
	(0.034)
Ex colonizer/colony	1.005***
	(0.043)
Common colonizer	0.615***
	(0.018)
Common religion	0.298***
	(0.025)
Common legal system before 1991	0.517***
	(0.018)
Common legal system after 1991	-0.471***
	(0.018)
Years at war	-0.067***
	(0.019)
Observations	202,709
R^2	0.736
Adjusted R^2	0.735
Residual Std. Error	2.045 (df = 202336)
F Statistic	1,512.800*** (df = 372; 202336)

Note: Data are from Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014) and Sarkees, Meredith Reid and Frank Wayman. (2010).
Standard errors are reported in the parentheses. *p **p ***p<0.01

5.2 Extension Results and Analysis

Having confirmed that our dataset and model is sufficiently similar to that of Melitz and Toubal (2014)'s, we move on to expanding the dataset. We had 202,709 observations in the replication dataset, which was more than doubled in the expanded version, which includes 516,013 observations. This is due to the observed years increasing from 1998-2007 to 1996-2020. Similarly to the replication section, the extension also includes three regressions; one with only LP_1, one with LP_2, and one with both LP variables. *Appendix 3* shows the regression with both legal system variables, separately from the regressions shown in Table 3. This is because the previously discussed issue with the Common legal system variable prevails in the extended dataset as well. The aim of this section is to compare the results from the larger dataset with our replication regression.

In the extended version with only one legal variable (Common legal system post 1991*)*, we see that all coefficients are quite similar to those of the replication version. (*see Table 3*) The signs, significance levels and magnitudes of the coefficients are similar to that of their counterpart in our replication regression. We see that among the language variables, the change in CSL and LP_1 is so miniscule it is inconsequential. COL and CNL do increase slightly, with an average of 0.046 and 0.059 increase. The change in LP_2 is noteworthy, in that in our replication regression, LP_2 was no longer significant when the model also includes LP_1, but this is not the case in the extension regression. LP_2 is still significant at our chosen significance level (with a higher coefficient as well; 0.028 as opposed to 0.019 in the replication), even when LP_1 is included in the regression. This might suggest that with more data at our disposal, LP_1 does not fully cover LP_2 as we thought based on the original, and replication regressions. Among the control variables, the only noteworthy change is that Distance, Common legal system and Years at war have coefficients of larger magnitude, whereas the other control variables decrease in terms of magnitude. The most important finding however, is that all variables (save for LP_2) remain on the same significance level as they had in the replication regression.

With a dataset that is 2.5 times larger than the original, the model produces similar results meaning that the original findings are more likely to be reliable, confirming our hypothesis. It is also evident from the results that Melitz and Toubal (2014)'s original hypothesis and regression, specifying what aspects of language are statistically important for bilateral trade, holds true. Not only do their language variables turn out to be significant, but the model has a high goodness-of-fit, and it holds up for a larger dataset.

Table 3: Common Language and Bilateral Trade: Extension Regression

	Dependent variable:		
	Bilateral trade (log)		
	(1)	(2)	(3)
Common official language	0.428***	0.442***	0.433***
	(0.015)	(0.015)	(0.015)
Common spoken language	0.463***	0.418***	0.447***
	(0.026)	(0.026)	(0.026)
Common native language	0.337***	0.399***	0.369***
	(0.035)	(0.037)	(0.037)
Linguistic proximity (1)	0.137***		0.122***
	(0.004)		(0.006)
Linguistic proximity (2)		0.169***	0.028***
		(0.006)	(0.009)
Distance (log)	-1.497***	-1.496***	-1.495***
	(0.005)	(0.005)	(0.005)
Common border	0.719***	0.734***	0.719***
	(0.022)	(0.022)	(0.022)
Ex colonizer/colony	0.965***	0.931***	0.963***
	(0.028)	(0.028)	(0.028)
Common colonizer	0.694***	0.679***	0.692***
	(0.012)	(0.012)	(0.012)
Common religion	0.224***	0.264***	0.223***
	(0.016)	(0.016)	(0.016)
Common legal system	-0.119***	-0.126***	-0.121***
	(0.008)	(0.008)	(0.008)
Years at war	-0.089***	-0.088***	-0.088***
	(0.013)	(0.013)	(0.013)
Observations	516,013	516,013	516,013
R^2	0.735	0.734	0.735
Adjusted R^2	0.734	0.734	0.734
Residual Std. Error	2.101 (df = 515627)	2.102 (df = 515627)	2.101 (df = 515626)
F Statistic	3,708.161*** (df = 385; 515627)	3,704.372*** (df = 385; 515627)	3,698.644*** (df = 386; 515626)

Note: Data are from Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014) and Sarkees, Meredith Reid and Frank Wayman. (2010).
Standard errors are reported in the parentheses. *p **p ***p<0.01

5.3 Integration and Examination of Other Variables

5.3.1. Exploring Endogeneity

Melitz and Toubal (2014) exclude all variables that may be affected by bilateral trade itself (i.e. are endogenous). However, excluding these may in some cases lead to omitted variable bias, for example if FTAs also have an effect on trade, independently of language and all other variables included in our regression. Their paper does include a section where they include some variables they deem to be exogenous, namely common currency, FTAs and cross-migration; the results show that only cross-migration is significant. We also replicated this regression on our extended dataset, but without cross-migration and adding GDP instead of common currency since many other works in this field do include it as an endogeneity check. We take the data for GDP - current thousands USD, unilateral - and FTAs - 1 if the pair is currently engaged in regional trade agreement, bilateral - from the CEPII database. We conducted our test by simply adding the FTA and GDP of the importer and exporter as additional variables in our regression. The new regression is thus specified in the following way:

$$log\ T_{i,j} = \beta_0 + \delta_c + \beta_1 COL + \beta_2 CSL + \beta_3 CNL + \beta_4 LP_{1,2} + \beta_5\ log\ D_{ij} + \beta_6 Cont_{ij} + \beta_7 ExCol_{ij} + \beta_8 ComCol_{ij} + \beta_9 ComRel_{ij} + \beta_{10} ComLeg_{ij} + \beta_{11} YearsAtWar_{ij} + \beta_{12} GDP_{origin} + \beta_{13} GDP_{destination} + \beta_{14} FTA_{ij} + \varepsilon_{ij}$$

The results are displayed in Table 4 below. The results are largely similar to our previous regressions with the extended dataset, in regards to the original variables used. The added variables GDP of the origin country, GDP of the destination country and whether they have a FTA prove to be significant at our chosen confidence level. However it is notable that both GDP variables are significant, but only have coefficients at 0.000, which is miniscule. This might indicate that while GDP is important to the level of trade, the other variables we have included, that often signify a preference for trade partners, sufficiently cover the variation in the level of bilateral trade. FTA on the other hand has a coefficient of 0.355, indicating that it does belong in the regression and intuitively has a positive and significant impact on trade. The change in the other variables is inadmissible. However, a further improvement would be taking the instrumental variable approach.

Table 4: Common Language and Bilateral Trade: Including Endogenous Variables

	Dependent variable:
	Bilateral trade (log)
Common official language	0.434***
	(0.015)
Common spoken language	0.278***
	(0.027)
Common native language	0.532***
	(0.037)
Linguistic proximity (1)	0.117***
	(0.006)
Linguistic proximity (2)	0.036***
	(0.009)
Distance (log)	-1.435***
	(0.005)
Common border	0.651***
	(0.023)
Ex colonizer/colony	1.007***
	(0.028)
Common colonizer	0.721***
	(0.012)
Common religion	0.178***
	(0.016)
Common legal system	-0.106***
	(0.008)
Years at war	-0.092***
	(0.013)
GDP (origin)	0.000***
	(0.000)
GDP (destination)	0.000***
	(0.000)
FTA	0.355***
	(0.011)
Observations	486,658
R^2	0.743
Adjusted R^2	0.743
Residual Std. Error	2.066 (df = 486272)
F Statistic	3,650.615*** (df = 385; 486272)

Note: Data are from Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014) and Sarkees, Meredith Reid and Frank Wayman. (2010).
Standard errors are reported in the parentheses. *p**p***p<0.01

In addition to exploring endogeneity with the addition of different variables, we also conducted other robustness checks. We tested for multicollinearity using the Variance Inflation Factor (VIF) test (see *Appendix 4.1*), analyzed our residual plot (see *Appendix 4.2*), and explored heteroscedasticity in our model with the Breusch-Pagan test (see *Appendix 4.3*). Overall, we found that the multicollinearity and

residual tests provide support for our model, but that there may be issues with heteroscedasticity that are important to consider in our analysis. See *Appendix 4* for more detail.

5.3.2 The Impact of Internet

As detailed earlier in our paper, we decided to investigate whether internet access has an effect on bilateral trade, on its own and more importantly through altering the importance of language. The first regression we run is the same linear model we have used thus far, run on our extended dataset, but controlling for internet access of the importer and the exporter country. The results are shown in the table in *Appendix 6*. (The results from the same model run on data from 2007 to 2020 can be found in *Appendix 5*.).

When comparing the same regression on the extended dataset without controlling for internet access, we see that most variables change slightly, but not in any observable systematic way. COL, CSL, CNL and LP_1 remain largely unchanged, with mild variation. However, LP_2 now has a negative coefficient, with a magnitude of -0.002. The internet variables themselves are significant at our chosen confidence level, indicating that access to the internet does influence bilateral trade. The magnitude of that influence however, is quite small; the coefficient for Internet of the exporter is 0.005, whereas Internet of the importer is -0.003.

As the relationship between Internet access, language similarity, and bilateral trade may be more complex, we composed interaction terms for all language variables with both variables indicating access to the internet. This will indicate whether Internet access acts as a moderator in the relationship between language similarity and bilateral trade. The regression is now specified as follows:

$$log\ T_{i,j} = \beta_0 + \delta_c + \beta_1 COL + \beta_2[COL * IntAcc_M] + \beta_3[COL * IntAcc_X] + \beta_4 CSL + \beta_5[CSL * IntAcc_M] + \beta_6[CSL * IntAcc_X] + \beta_7 CNL + \beta_8[CNL * IntAcc_M] + \beta_9[CNL * IntAcc_X] + \beta_{10} LP_{1,2} + \beta_{11}(LP_{1,2}[LP_{1,2} * IntAcc_M]) + \beta_{12}(LP_{1,2}[LP_{1,2} * IntAcc_X]) + \beta_{13} IntAcc_X + \beta_{14} IntAcc_M + \beta_{15} ExCol_{ij} + \beta_{16} ComCol_{ij} + \beta_{17} ComRel_{ij} + \beta_{18} ComLeg_{ij} + \beta_{19} YearsAtWar_{ij} + \varepsilon_{ij}$$

where M stands for importer country and X stands for exporter country.

The results displayed in Table 5, found below, are quite telling. We see that the interaction terms between the internet access of the exporter and almost all language variables with the exception of COL, are significant at our chosen confidence level, and importantly, the coefficients are negative for all interaction terms except for between LP_2 and CNL, and the internet variables. This implies that a

positive change of unit in internet access of the exporter country actually decreases the magnitude of the relationship between bilateral trade and CSL, and LP_1, in accordance with our hypothesis.

Table 5: Common Language and Bilateral Trade: The Impact of Internet

	Dependent variable:
	Bilateral trade (log)
Common official language	0.363***
	(0.021)
COL*Importer Internet	-0.003***
	(0.0004)
COL*Exporter Internet	0.001*
	(0.0004)
Common spoken language	1.548***
	(0.045)
CSL*Importer Internet	-0.009***
	(0.001)
CSL*Exporter Internet	-0.014***
	(0.001)
Common native language	-0.209***
	(0.060)
CNL*Importer Internet	0.001
	(0.001)
CNL*Exporter Internet	0.012***
	(0.001)
Language proximity (1)	0.269***
	(0.009)
LP1*Importer Internet	-0.0002
	(0.0002)
LP1*Exporter Internet	-0.004***
	(0.0002)
Language proximity (2)	-0.098***
	(0.015)
LP2*Importer Internet	0.001**
	(0.0003)
LP2*Exporter Internet	0.003***
	(0.0003)
Internet access (exporter)	0.008***
	(0.0003)
Internet access (importer)	-0.001*
	(0.0003)
Observations	483,476
R^2	0.743
Adjusted R^2	0.742
Residual Std. Error	2.071 (df = 483080)
F Statistic	3,527.791*** (df = 395; 483080)

Note: Data are from Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014), Sarkees, Meredith Reid and Frank Wayman. (2010), and World Bank. (2023). Standard errors are reported in the parentheses. *p **p ***p<0.01

The interaction between the internet access of the importer country and CSL is also significant and negative, as with COL. Interestingly, LP_2 and CNL and internet have positive coefficients, meaning that an increase in access to the internet also increases the effect linguistic similarity has on bilateral trade. This is counterintuitive, as we did not expect to see that having access to the internet would make sharing a native language more important. However, this might be that due to the addition of the internet, CNL and LP_2 take on a role of cultural variables, representing emotional and cultural reasons for choosing trade partners, instead of ease of communication. Overall, access to the Internet seems to lower the impact of common language on bilateral trade, which is in line with our initial hypothesis. However, our interpretation is that when controlling for the internet, only those language variables that represent ease of communication will be negatively affected, as access to the internet smoothes out trade barriers that stem from language, but does not affect any cultural affiliations any two countries might have.

5.3.3 Using Updated Language Data

Next, we run our regressions using Toubal's more recent language data, as described in the literature review. This means that we substitute the original "COL" and "CNL" in all the columns. However, Toubal's new dataset did not contain a new "CSL" variable, so we decided to continue using the original for this specific variable. Similarly, Toubal constructed only one variable for linguistic proximity, which is constructed closer to the method of the original LP_1 variable. For these reasons, we ran three separate regressions, shown in Table 6, which contains 3 columns; one with the original LP_1 variable (Table 6, column 1), one with the original LP_2 variable (Table 6, column 2), and a final regression with the new LP variable (Table 6, column 3).

When comparing these results with the corresponding one in our regressions run on the extended dataset, we see that some variables change drastically. The most notable change is the newly constructed Language proximity variable, which is significant with a high coefficient of 1.091. Other notable differences when LP is used in language coefficients are following; COL decreases, CSL increases highly, while CNL turns strongly negative, decreasing from 0.369 (see *Table 3, column 3*) to -1.335, a drastic change. We have no explanation for this, as it would suggest that with better LP data, common native language actually turns out to be a strong deterrent of trade.

The control variables change slightly, with the colonizer variables and common legal system increasing marginally, common border and years at war decreasing slightly, and distance and common religion staying virtually unchanged.

Table 6: Common Language and Bilateral Trade: Extension Regression with updated language data

	Dependent variable:		
	Bilateral trade (log)		
	(1)	(2)	(3)
Common official language	0.394***	0.402***	0.362***
	(0.012)	(0.012)	(0.012)
Common spoken language	0.877***	0.886***	0.641***
	(0.022)	(0.022)	(0.023)
Common native language	-0.613***	-0.701***	-1.335***
	(0.040)	(0.040)	(0.041)
Linguistic proximity (1)	0.099***		
	(0.004)		
Linguistic proximity (2)		0.103***	
		(0.005)	
Linguistic proximity			1.091***
			(0.026)
Distance (log)	-1.500***	-1.503***	-1.465***
	(0.005)	(0.005)	(0.005)
Common border	0.648***	0.658***	0.586***
	(0.022)	(0.022)	(0.023)
Ex colonizer/colony	0.990***	0.967***	1.021***
	(0.028)	(0.028)	(0.029)
Common colonizer	0.698***	0.689***	0.708***
	(0.012)	(0.012)	(0.012)
Common religion	0.239***	0.278***	0.229***
	(0.016)	(0.016)	(0.016)
Common legal system	-0.150***	-0.156***	-0.158***
	(0.008)	(0.008)	(0.008)
Years at war	-0.093***	-0.094***	-0.098***
	(0.013)	(0.013)	(0.013)
Observations	516,014	516,014	534,622
R^2	0.735	0.735	0.738
Adjusted R^2	0.735	0.734	0.738
Residual Std. Error	2.101 (df = 515628)	2.101 (df = 515628)	2.142 (df = 534236)
F Statistic	3,711.247*** (df = 385; 515628)	3,708.096*** (df = 385; 515628)	3,906.690*** (df = 385; 534236)

Note: Data are from Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014), Sarkees, Meredith Reid and Frank Wayman. (2010), and Gurevich, Tamara, Peter Herman, Farid Toubal, and Yoto Yotov, (2021).
Standard errors are reported in the parentheses. *p **p ***p<0.01

5.4 Evolvement Over Time

As we see some differences, mainly in magnitude of coefficients, between the replication and extension results, we decided to investigate how the results differ depending on the time period. Specifically, we find it pertinent to compare the time period before and after 2007, which is the year in which Melitz and Toubal (2014) end their research. To clarify, we now return to using the same variables as in *Table 1*. Additionally, 2007 can be considered a turning point for the use of electronics etc. with standards changing in both developed and developing countries (Naughton. 2016). 2007 was the year the first iPhone was released, marking the beginning of a new era. According to our hypothesis, technology and the availability of information and language resources should affect the language variables negatively, i.e. decrease the magnitude of the language coefficients.

It is prudent to compare the results with the replication regression, since the datasets are closer in size - this dataset (from 2007-2020) contains 309,367 observations, as opposed to the replication data containing 202,709 (1996-2007). *Appendix 7* shows our findings for the time period from 2007 to 2020. We see that the results do not support our hypothesis that increased technological access will decrease the importance of language variables, through the introduction of smartphones on the global market. All variables have the same significance level as in the dataset run on the years 1996 to 2007. Notably, some of the language variables have a higher coefficient than in the replication regression. The difference might partly be due to the data being of better quality than in earlier years, where data gathering posed a problem. We see that adjusted R-squared at 74.5% at is also higher than in previous regressions, supporting that theory. However, the results do show that all language variables are all still significant in the examined time frame. We suggest more potential reasons for the variation in magnitude in the different time periods with the subsequent analysis.

More importantly, we ran six separate regressions on different year subsets. The purpose of this was twofold - firstly, it allows us to conduct a more detailed temporal analysis, identify trends or pattern shifts overtime, and comment on future expectations. Secondly, we can make further inferences regarding the model fit. Below are our results for the different years, in Table 7 and Table 8. Note that we decided to select every five years from the beginning of our data, beginning from 1996 and ending in 2020.

Table 7: Common Language and Bilateral Trade: 1996, 2001, 2006

	Dependent variable:		
	Bilateral trade (log)		
	1996	2001	2006
	(1)	(2)	(3)
Common official language	0.102	0.211***	0.372***
	(0.085)	(0.073)	(0.070)
Common spoken language	0.694***	0.481***	0.336***
	(0.140)	(0.128)	(0.124)
Common native language	-0.442**	0.342*	0.464***
	(0.205)	(0.179)	(0.174)
Linguistic proximity (1)	-0.075**	0.111***	0.169***
	(0.035)	(0.031)	(0.029)
Linguistic proximity (2)	0.204***	0.022	-0.020
	(0.049)	(0.045)	(0.042)
Distance (log)	-1.261***	-1.469***	-1.512***
	(0.029)	(0.024)	(0.023)
Common border	0.243*	0.596***	0.589***
	(0.128)	(0.108)	(0.106)
Ex colonizer/colony	1.279***	1.074***	0.988***
	(0.118)	(0.136)	(0.136)
Common colonizer	0.611***	0.673***	0.685***
	(0.073)	(0.059)	(0.055)
Common religion	0.356***	0.254***	0.203***
	(0.094)	(0.082)	(0.075)
Common legal system	-0.352***	-0.541***	-0.497***
	(0.066)	(0.056)	(0.054)
Years at war	-0.029	-0.069	-0.107*
	(0.053)	(0.062)	(0.062)
Observations	11,328	19,961	22,076
R^2	0.787	0.737	0.752
Adjusted R^2	0.780	0.732	0.748
Residual Std. Error	1.673 (df = 10964)	2.051 (df = 19597)	2.064 (df = 21712)
F Statistic	111.453*** (df = 363; 10964)	150.907*** (df = 363; 19597)	181.058*** (df = 363; 21712)

Note: Data are from Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014) and Sarkees, Meredith Reid and Frank Wayman. (2010).
Standard errors are reported in the parentheses. *p **p ***p<0.01

Table 8: Common Language and Bilateral Trade: 2011, 2016, 2020

	Dependent variable:		
	Bilateral trade (log)		
	2011	2016	2020
	(1)	(2)	(3)
Common official language	0.431***	0.451***	0.331***
	(0.071)	(0.068)	(0.071)
Common spoken language	0.395***	0.419***	0.491***
	(0.128)	(0.123)	(0.127)
Common native language	0.462***	0.627***	0.513***
	(0.178)	(0.173)	(0.179)
Linguistic proximity (1)	0.160***	0.115***	0.117***
	(0.030)	(0.029)	(0.030)
Linguistic proximity (2)	-0.019	0.058	0.012
	(0.045)	(0.043)	(0.044)
Distance (log)	-1.563***	-1.505***	-1.420***
	(0.023)	(0.023)	(0.024)
Common border	0.603***	0.728***	0.801***
	(0.110)	(0.106)	(0.111)
Ex colonizer/colony	0.925***	0.841***	0.901***
	(0.140)	(0.135)	(0.134)
Common colonizer	0.669***	0.672***	0.736***
	(0.056)	(0.054)	(0.057)
Common religion	0.103	0.019	0.052
	(0.077)	(0.074)	(0.077)
Common legal system	-0.472***	-0.490***	-0.438***
	(0.055)	(0.053)	(0.053)
Years at war	-0.116*	-0.119*	-0.112*
	(0.064)	(0.061)	(0.061)
Observations	22,226	22,269	20,617
R^2	0.748	0.761	0.768
Adjusted R^2	0.744	0.757	0.764
Residual Std. Error	2.117 (df = 21864)	2.040 (df = 21907)	2.017 (df = 20255)
F Statistic	180.033*** (df = 361; 21864)	193.169*** (df = 361; 21907)	185.419*** (df = 361; 20255)

Note: Data are from Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014) and Sarkees, Meredith Reid and Frank Wayman. (2010).
Standard errors are reported in the parentheses. *p **p ***p<0.01

Firstly, we note that the number of observations increases every five years. In 1996, we only had 11,328 observations, while in 2001 we increased to 19,961. This pattern persists for each regression except the final year 2020, where the number of observations falls slightly. This is important to keep in mind as the different number of observations across subsets may lead to interpretation challenges, statistical power changes, and bias. For example, our results from 1996 show that COL is not statistically

significant while it is for all other year samples, and the magnitude of the other language coefficients (CSL = 0.694, CNL= -0.442, LP_1 = -0.075, LP_2 = 0.204) are very different from the other years. We thus remain cautious with our comparison of 1996 with the other years. Looking at the rest of the years, we see that our results for COL, CSL, CNL, and LP_1 are all significant at a 1% significance level with the exception of 2001 where CNL is only significant at a 5% level. All coefficients of LP_2 are statistically insignificant, which aligns with our previous results where all language variables are run together. Looking at each language variable separately, we see a pattern of growth in magnitude of the COL coefficient up to 2016, suggesting that the institutional language is increasingly important in trade relations. Inspecting CSL, we see that in 1996 the coefficient is much larger than what we have seen in previous regressions (0.694), decreases in magnitude up till 2006, and then grows for the remaining period. This initial decreasing pattern is likely due to improved communication and information technologies in the early 21st century, reducing communication costs between countries. It is difficult to firmly deduce the reason for the subsequent increase in importance of spoken language, but it could be due to changing global trade patterns impacted by for instance the financial crisis of 2008. Similarly, CNL has an overall increasing trend, indicating that native language has a larger impact on trade than it had in the past. This may suggest that cultural ties and local markets are perhaps becoming more important, such as the emphasis of production and consumption of local goods. Finally, LP_1 stays approximately the same (around 0.1 and 0.2) throughout 2001-2020. This result is intuitive as the familial roots of languages do not change overtime, and thus our results show that the impact of this aspect of language on bilateral trade also remains relatively constant overtime.

6. Conclusion and discussion

Our results show that Melitz and Toubal (2014)'s findings - that language has a much higher impact on bilateral trade and is more nuanced than previously thought - stand firm when employed on a dataset that is double the size of their original dataset. This is also verified through many robustness checks, although heteroscedasticity seems to be an issue. Our interpretation is that this is due to some countries being clustered based on potential aspects we have not accounted for, such as developing versus developed countries, or clusters based on distance. This has potential for future investigation.

A general observation about the language variables is that including LP_1 instead of LP_2 increases the effect of Common official language, Common native language and colonial ties in terms of the magnitude of their coefficients (including LP_2 has the opposite effect). It is likely that the reason for this is the conceptual difference between the historical and ethnic aspect of language, which LP_1 and CNL represent, versus the communication aspect that LP_2 and CSL represent. However, we see no advantage in using either one instead of the other in future research.

Another notable finding is that with improved language data, Common native language turns negative, which is a highly surprising development. We are confident in the validity of our regression, but we cannot give a certain explanation for this. Common native language is usually considered to be an influential stand-in for common culture, which is widely accepted to be a facilitator of bilateral trade. Future research should focus on investigating common native language more closely.

One of our most important findings is that in accordance with our hypothesis, access to the Internet does in fact lower the impact common language has on trade. This effect is stronger when looking at the exporter country, as opposed to the country importing goods. We believe that by being able to easily access information about languages, as well as translation tools, the role of official translators decreases, as well as those aspects of ease of communication that previously came from common language. The full extent of the interaction between Internet access, common language, and even information technology is not covered in our paper, and requires further investigation. It is warranted however by our findings. The importance of finding out how impactful internet access is is easy to see. If the evolution of information technology up until 2020 influences the relationship between language and bilateral trade, one can assume that the continued exponential growth and spread of internet access will continue to diminish the importance of language. This might mean that policy makers have a choice; should native languages be preserved for the sake of preservation, or be abandoned, due to its economic significance decreasing? Here it is important to note, however, that when we tested the evolution of the language variables overtime, we saw an overall positive trend in the coefficients for COL, CSL, and CNL, which seemingly counteracts the intuition. It may be due to insufficient controls for factors important in the later 2000s, as we see a different pattern when for example internet is included.

During our research, we also identified additional areas of interest that could be investigated in future research. For instance, it could be fruitful to compare the impact of language on different types of trade. Like Melitz' and Toubal's investigation into the different classifications of goods (Rauch classification) it would be interesting to research whether there is a difference in the importance of language when trading products versus services. It seems intuitive that language would play a more central role in the exchange of services due to the implicit requirement of communication in the service sector. Another area that could benefit from more extensive research is the distinction between direct and indirect communication which Melitz, Toubal and our research touches upon. This especially can have implications for policy-making and the importance of translation and interpreters. For example, language learning in schools and the language of academia could be impacted; if machine translation proves insubstantial in a more extensive experiment, it shows that language learning should not be deprioritized. In conclusion, society has much to learn from further research in this topic, which is bound to become more relevant in our increasingly global economy.

7. References

Anderson, J. and van Wincoop, E. (2003) *"Gravity with gravitas: A solution to the border puzzle"*, The American Economic Review , Mar., 2003, Vol. 93, No. 1 (Mar., 2003), pp. 170- 192

Baier, S., & Standaert, S. (2020, March 31). "*Gravity Models and Empirical Trade"*. Oxford Research Encyclopedia of Economics and Finance. Retrieved 15 Sep. 2023, from https://oxfordre.com/economics/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-327.

Brynjolfsson, E., Hui, X. and Liu, M. (2018) "*Does machine translation affect international trade? evidence from a large digital platform*," *NBER.* Available at: https://www.nber.org/papers/w24917 (Accessed: 03 December 2023).

Cheng, I.-H. (2005) "*Controlling for heterogeneity in gravity models of trade and Integration"*. Available at: https://files.stlouisfed.org/files/htdocs/publications/review/05/01/Cheng.pdf (Accessed: 03 December 2023).

Coe, D.T. *et al.* (2002) "*The missing globalization puzzle" - WP/02/171 - IMF*. Available at: https://www.imf.org/external/pubs/ft/wp/2002/wp02171.pdf (Accessed: 03 December 2023).

Deardorff, A.V. (1998) "*Determinants of bilateral trade: Does gravity work in a neoclassical world?"* Available at: https://www.nber.org/system/files/chapters/c7818/c7818.pdf (Accessed: 03 December 2023).

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2023. "*Ethnologue: Languages of the World".* Twenty-sixth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

Egger, P.H. and Lassmann, A. (2012) "*The language effect in International Trade: A meta-analysis*", Economics Letters, 116(2), pp. 221–224. doi:10.1016/j.econlet.2012.02.018.

Fensore, I., Legge, S. and Schmid, L. (2022) "*Ancestry and International Trade*", Journal of Comparative Economics, 50(1), pp. 33–51. doi:10.1016/j.jce.2021.05.002.

Fidrmuc, J., Fidrmuc, J. (2016). "*Foreign languages and trade: evidence from a natural experiment"*. Empir Econ **50**, 31–49 https://doi.org/10.1007/s00181-015-0999-7

Foroutan, F. (1992) "*Regional Integration in Sub-Saharan Africa*" Policy Research Dissemination Center, The World Bank (WPS992), Regional Integration in Sub-Saharan Africa - Faezeh Foroutan - Google-kirjat

Frankel, J., Stein, E. and Wei, S. (1993) "*Continental Trading Blocs: Are they natural or super-natural?"* NBER working paper series, Available at: https://www.nber.org/system/files/working_papers/w4588/w4588.pdf (Accessed: 03 December 2023).

Gurevich, T. et al., (2021) "*One nation, one language? domestic language diversity, trade and Welfare"*, SSRN Electronic Journal [Preprint]. doi:10.2139/ssrn.3774667.

Havrylyshyn, O. and Pritchett, L. (1991) "*European trade patterns after the transition"* - documents1.worldbank.org. Available at: https://documents1.worldbank.org/curated/en/891101468751525610/pdf/multi0page.pdf (Accessed: 03 December 2023).

Head, K. and Mayer, T. (2013) "*Gravity equations: Workhorse, toolkit and Cookbook - Cepii, CEPII"*. Available at: http://www.cepii.fr/PDF_PUB/wp/2013/wp2013-27.pdf (Accessed: 03 December 2023).

"International Decade of Indigenous Languages 2022 – 2032 for Indigenous Peoples (no date) United Nations". Available at: https://www.un.org/development/desa/indigenouspeoples/indigenous-languages.html (Accessed: 17 October 2023).

Head, K., Mayer, T., Ries J.,(2011), "*The erosion of colonial trade linkages after independence." Journal of International Economics,* 81 (1), pp.1-14. ff10.1016/j.jinteco.2010.01.002ff. ffhal01024396f

Kitenge, E., and Lahiri, S. (2021). "*Is the Internet bringing down language-based barriers to international trade?"*. Review of International Economics. 30. 10.1111/roie.12576.

"LANGUAGE FAMILY" . Concise Oxford Companion to the English Language. Encyclopedia.com. (September 18, 2023). https://www.encyclopedia.com/humanities/encyclopedias-almanacs-transcripts-and-maps/language-family

Linnemann, H. (1966) "*An econometric study of International Trade Flows"*. Amsterdam: North-Holland Pub.

Lohmann, J. (2011) "*Do language barriers affect trade?"*, Economics Letters, 110(2), pp. 159–162. doi:10.1016/j.econlet.2010.10.023.

Mahfuz Kabir, Ruhul Salim, Nasser Al-Mawali, (2017), "*The gravity model and trade flows: Recent developments in econometric modeling and empirical evidence",* Economic Analysis and Policy, Vol. 56, 60-71, ISSN 0313-5926, https://doi.org/10.1016/j.eap.2017.08.005.

Manacorda, M. and Tesei, A. (2020), "*Liberation Technology: Mobile Phones and Political Mobilization in Africa"*. Econometrica, 88: 533-567. https://doi.org/10.3982/ECTA14392

Mayer, T. and Zignano, S. (2011) "*Notes on Cepii's distances measures: The geodist database*." Available at: http://www.cepii.fr/PDF_PUB/wp/2011/wp2011-25.pdf (Accessed: 03 December 2023).

Melitz and Toubal (2018) "*Somatic distance, trust and trade - cepii, CEPII"*. Available at: http://www.cepii.fr/PDF_PUB/wp/2018/wp2018-11.pdf (Accessed: 03 December 2023).

Melitz and Toubal (2014) *"Native language, spoken language, translation and Trade",* Journal of International Economics, 93(2), pp. 351–363. doi:10.1016/j.jinteco.2014.04.004.

Melitz, J. (2007) "*Language and foreign trade, European Economic Review"*. Available at: https://www.sciencedirect.com/science/article/pii/S0014292107000621 (Accessed: 03 December 2023).

Naughton, J. (2016) "*2007, not 2016, is the year the world turned upside down", The Guardian*. Available at: https://www.theguardian.com/commentisfree/2016/nov/27/2007-not-2016-year-world-turned-upside-down-rapid-technological-change (Accessed: 04 December 2023).

Ribeiro, S. and Ferro, M. (n.d.). "*The Influence of Language Similarity in International Trade: Evidence from Portuguese Exports in 2013"*. [online] Available at: https://research.unl.pt/ws/portalfiles/portal/3325759/The_influence_of_language_similarity_in_international_trade.pdf

Rose, A. K., Lockwood, B., & Quah, D. (2000). "*One Money, One Market: The Effect of Common Currencies on Trade".* Economic Policy, *15*(30), 9–45. http://www.jstor.org/stable/1344722

Toubal, F. and Melitz, J. (2019) "*The potential impact of machine translation on foreign trade – caution, please", CEPR*. Available at: https://cepr.org/voxeu/columns/potential-impact-machine-translation-foreign-trade-caution-please (Accessed: 03 December 2023).

Zhang, Weiguo, and Gilles Grenier., 2012., "*How can Language be linked to Economics? A Survey of Two Strands of Research".* Faculty of Social Sciences. https://socialsciences.uottawa.ca/economics/sites/socialsciences.uottawa.ca.economics/files/1206E.pdf.

7.1 Tables and Data Sources

Tables: Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022). "*The CEPII Gravity Database* [Data set]," CEPII Working Paper 2022- 05 , July 2022 , CEPII. CEPII - The CEPII Gravity Database

Melitz and Toubal (2014). "*Native Language, Spoken Language, Translation and Trade* [Data set],". Journal of International Economics, Vol. 92, N°2: 351-363. CEPII - Language

Sarkees, Meredith Reid and Frank Wayman. (2010). "*Resort to War: 1816 – 2007* [Data set]," Washington DC: CQ Press. COW War Data, 1816 – 2007 (v4.0) – Correlates of War

Gurevich, T., Herman, P., Toubal, F. and Yokov, Y. (2021) "*Gravity Portal: DICL, Gravity Portal: DICL* [Data set] | *United States International Trade Commission"*. https://www.usitc.gov/data/gravity/dicl.htm

World Bank. (2023). "*Individuals using the Internet (% of population)* [Data set]", International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database, CC BY-4.0. Individuals using the Internet (% of population) | Data (worldbank.org)

8. Appendix

Appendix 1.

Melitz and Toubal (2014) general results (retrieved from Melitz and Toubal (2014)*, page 42)*

Table 3: Common language
Regressand: log of bilateral trade (Total)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Common official language	0.514				0.316	0.360	0.351	0.431
	(13.518)				(6.864)	(7.716)	(7.561)	(9.740)
Common spoken language		0.775			0.503	0.399	0.396	
		(14.651)			(6.578)	(5.104)	(4.910)	
Common native language			0.856		0.062	0.294	0.284	0.639
			(11.227)		(0.573)	(2.588)	(2.344)	(6.755)
Common native language dummy				0.684				
				(11.568)				
Linguistic proximity (tree)						0.073		
						(6.170)		
Linguistic proximity (ASJP)							0.078	0.105
							(4.253)	(6.048)
Distance (log)	-1.394	-1.379	-1.385	-1.386	-1.375	-1.364	-1.365	-1.366
	(-90.272)	(-87.949)	(-88.075)	(-87.982)	(-87.679)	(-86.392)	(-86.420)	(-86.458)
Common border	0.722	0.671	0.719	0.718	0.679	0.662	0.670	0.690
	(8.413)	(7.766)	(8.345)	(8.337)	(7.885)	(7.723)	(7.817)	(8.077)
Ex colonizer/colony	1.484	1.579	1.653	1.666	1.472	1.500	1.484	1.501
	(14.347)	(15.297)	(15.757)	(15.934)	(14.329)	(14.588)	(14.426)	(14.506)
Common colonizer	0.754	0.851	0.909	0.908	0.780	0.775	0.779	0.785
	(16.687)	(19.461)	(20.636)	(20.613)	(17.085)	(16.957)	(17.045)	(17.102)
Common religion	0.429	0.329	0.416	0.406	0.325	0.264	0.289	0.319
	(8.664)	(6.475)	(8.293)	(8.081)	(6.383)	(5.087)	(5.589)	(6.210)
Common legal system	0.244	0.311	0.274	0.278	0.240	0.209	0.217	0.189
	(6.817)	(9.029)	(7.695)	(7.825)	(6.544)	(5.666)	(5.866)	(5.202)
Years at war	-0.398	-0.417	-0.385	-0.389	-0.397	-0.382	-0.382	-0.365
	(-2.388)	(-2.501)	(-2.357)	(-2.391)	(-2.382)	(-2.272)	(-2.283)	(-2.188)
Observations	209276	209276	209276	209276	209276	209276	209276	209276
Adjusted R^2	0.756	0.756	0.756	0.756	0.757	0.757	0.757	0.757
Number of clusters	28950	28950	28950	28950	28950	28950	28950	28950

All regressions contain exporter/year and importer/year fixed effects. Student *t*s are in parentheses. These are based on robust standard errors that have been adjusted for clustering by country pair.

Appendix 2.

Running each linguistic variable on its own

Common Language and Bilateral Trade: Replication Regression Aggregate

	Dependent variable:		
	Bilateral trade (log)		
	(1)	(2)	(3)
Common official language	0.542***		
	(0.019)		
Common spoken language		0.811***	
		(0.028)	
Common native language			0.785***
			(0.039)
Distance (log)	-1.521***	-1.506***	-1.510***
	(0.007)	(0.007)	(0.007)
Common border	0.765***	0.729***	0.769***
	(0.035)	(0.035)	(0.035)
Ex colonizer/colony	0.997***	1.083***	1.133***
	(0.043)	(0.043)	(0.043)
Common colonizer	0.658***	0.756***	0.797***
	(0.018)	(0.017)	(0.017)
Common religion	0.562***	0.439***	0.549***
	(0.023)	(0.025)	(0.024)
Common legal system	-0.075***	-0.069***	-0.037***
	(0.012)	(0.012)	(0.012)
Years at war	-0.077***	-0.085***	-0.078***
	(0.020)	(0.020)	(0.020)
Observations	205,235	205,235	205,235
R^2	0.735	0.735	0.734
Adjusted R^2	0.734	0.734	0.734
Residual Std. Error (df = 204867)	2.076	2.076	2.078
F Statistic (df = 367; 204867)	1,547.066***	1,547.639***	1,543.404***

Note: Data are from *Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014) and Sarkees, Meredith Reid and Frank Wayman. (2010).*
Standard errors are reported in the parentheses. *p**p***p<0.01

Note: when all three are additionally ran together (not shown here), we also see what Column 5 of Appendix 1 shows, namely CNL becomes insignificant when ran together with COL and CSL

Appendix 3

Including all control variables, run on extended data

Common Language and Bilateral Trade: Extension Regression with all variables

	Dependent variable:
	Bilateral trade (log)
Common official language	0.388***
	(0.015)
Common spoken language	0.434***
	(0.026)
Common native language	0.369***
	(0.037)
Linguistic proximity (1)	0.126***
	(0.006)
Linguistic proximity (2)	0.011
	(0.009)
Distance (log)	-1.491***
	(0.005)
Common border	0.679***
	(0.022)
Ex colonizer/colony	0.945***
	(0.028)
Common colonizer	0.635***
	(0.012)
Common religion	0.218***
	(0.016)
Common legal system before 1991	0.487***
	(0.011)
Common legal system after 1991	-0.469***
	(0.011)
Years at war	-0.085***
	(0.013)
Observations	516,013
R^2	0.736
Adjusted R^2	0.735
Residual Std. Error	2.098 (df = 515625)
F Statistic	3,706.749*** (df = 387; 515625)

Note: Data are from *Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014) and Sarkees, Meredith Reid and Frank Wayman. (2010).*
Standard errors are reported in the parentheses. *p **p ***p<0.01

Appendix 4

Robustness checks

Appendix 4.1 Multicollinearity with VIF test

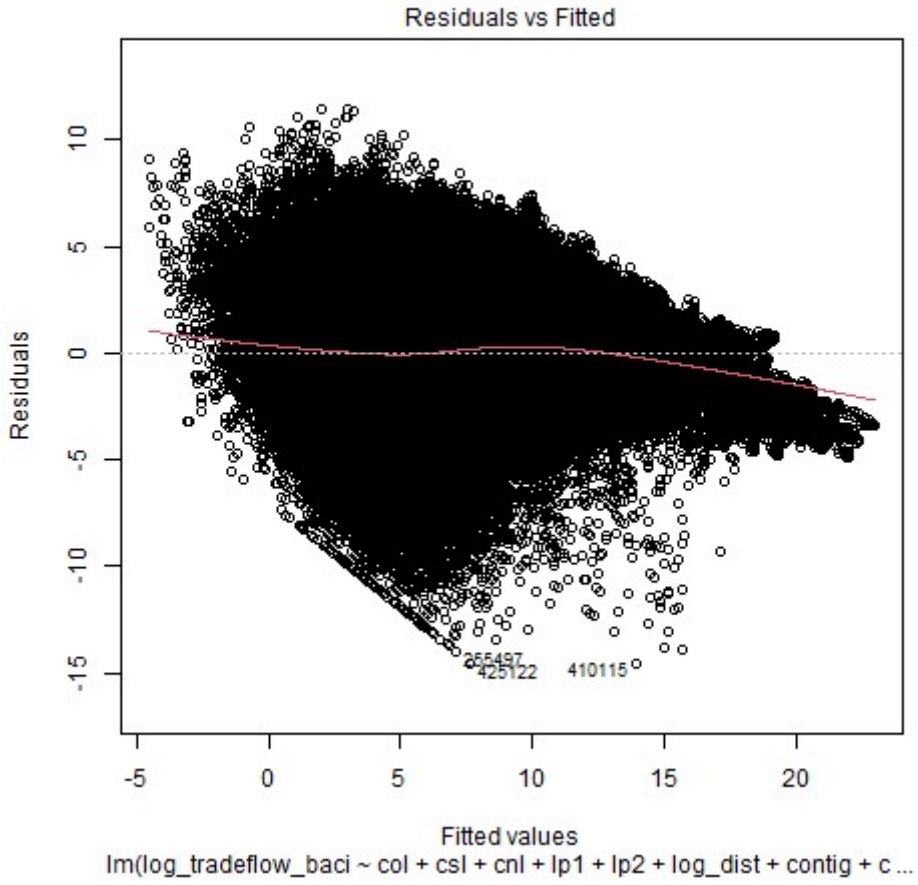
Firstly, we checked the data for multicollinearity, the method of which we choose to be a VIF test. The results of that test are shown below. We see that all variables are below a value of 5, meaning that those variables are not highly correlated with another variable. The exemption for this is the two Language proximity variables, both exceeding 5 at 5.98 and 5.79 respectively, which was expected. Logically linguistic proximity and linguistic similarity are quite highly correlated, since languages deriving from the same root are likely to be similar when spoken.

Variance Inflation Factor (VIF) Results

	GVIF	DF	GVIF^(1/(2*Df))
COL	2.972823	1	1.724188
CSL	4.421130	1	2.102648
CNL	3.593087	1	1.895544
LP1	5.987423	1	2.446921
LP2	5.792555	1	2.406773
Distance (log)	1.854415	1	1.361769
Common border	1.229156	1	1.108672
Ex colonizer/colony	1.312890	1	1.145814
Common colonizer	2.331738	1	1.527003
Common religion	1.865398	1	1.365796
Common legal system (pre-1991)	3.335350	1	1.826294
Common legal system (post-1991)	2.522680	1	1.879809
Years at war	1.084759	1	1.041518
Factor (iso3_o)	12.836483	175	1.007319
Factor (iso3_d)	12.519033	175	1.007247
Factor (year)	1.043740	24	1.000892

Appendix 4.2 Residual plot

Next, we did a scatterplot of the residual errors to see how they are distributed. We see that the residuals of the errors do not seem to have a logical order of distribution, but it is hard to determine with 100% confidence, since the errors are so clustered together. As shown on the graph, the fitted model is quite close to the observed reality and the residuals seem to center around zero, although it deviates more in the beginning and at the end of the examined time frame. This might be due to the fact there are significantly more observations towards the middle of the time frame, where, notably, the model fits the real observations quite well. There also might be unobserved influences on trade in the third part of the time frame, which indicates a need to explore further. We do this by adding internet access as a variable to our data, and by subsetting it to the period between 2007 and 2021 to see how the model behaves then.



Appendix 4.3 Breusch - Pagan test

studentized Breusch-Pagan test
data: extensionregressionall
BP = 60373, df = 387, p-value < 2.2e-16

Heteroscedasticity can also be a source of worry for the validity of our results. Unfortunately, the Breusch-Pagan test shows that the p-value for the test is not significant at our chosen significance level of 1%, meaning that the test cannot reject the null hypothesis that there is heteroscedasticity in the model. This might be due to omitted variable bias, misspecification of the model, or heterogeneous groups. The first two are difficult to deal with; we have chosen not to include more variables to stay as close to our benchmark paper as possible, and the model performs well in other tests, also producing very similar results to what Melitz and Toubal (2014) reached. Regarding heterogeneous groups, we hypothesize that there is some previously unaccounted for similarity between some groups of countries that we have not included in our model. A potential investigation could be conducted on whether grouping the countries based on distance would have an impact on the outcome. Country pairs that are further away from each other than the median value for Distance might behave more similarly to each other, than to country pairs that are closer than the median value for Distance.

Appendix 5

Impact of language on trade when controlling for internet access (2007-2020)

Common Language and Bilateral Trade: The Impact of Internet, 2007-2020

	Dependent variable:
	Bilateral trade (log)
Common official language	0.439***
	(0.019)
Common spoken language	0.349***
	(0.035)
Common native language	0.552***
	(0.049)
Linguistic proximity (1)	0.140***
	(0.008)
Linguistic proximity (2)	0.007
	(0.012)
Distance (log)	-1.536***
	(0.006)
Common border	0.660***
	(0.030)
Ex colonizer/colony	0.863***
	(0.038)
Common colonizer	0.642***
	(0.015)
Common religion	0.136***
	(0.021)
Common legal system before 1991	0.478***
	(0.015)
Common legal system after 1991	-0.474***
	(0.015)
Years at war	-0.113***
	(0.017)
Internet (Exporter)	0.001**
	(0.001)
Internet (Importer)	0.001
	(0.001)
Observations	290,604
R^2	0.751
Adjusted R^2	0.750
Residual Std. Error	2.082 (df = 290229)
F Statistic	2,336.578*** (df = 374; 290229)

Note: Data are from Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014), Sarkees, Meredith Reid and Frank Wayman. (2010), and World Bank. (2023). Standard errors are reported in the parentheses. *p**p***p<0.01

Appendix 6

Internet as a control variable in our extended regression.

Common Language and Bilateral Trade: Internet as control variable

	Dependent variable:
	Bilateral trade (log)
Common official language	0.389***
	(0.015)
Common spoken language	0.394***
	(0.027)
Common native language	0.479***
	(0.037)
Linguistic proximity (1)	0.137***
	(0.006)
Linguistic proximity (2)	-0.002
	(0.009)
Distance (log)	-1.494***
	(0.005)
Common border	0.645***
	(0.023)
Ex colonizer/colony	0.948***
	(0.028)
Common colonizer	0.629***
	(0.012)
Common religion	0.205***
	(0.016)
Common legal system before 1991	0.485***
	(0.012)
Common legal system after 1991	-0.460***
	(0.012)
Years at war	-0.091***
	(0.013)
Internet (Exporter)	0.005***
	(0.0003)
Internet (Importer)	-0.003***
	(0.0003)
Observations	483,476
R^2	0.741
Adjusted R^2	0.741
Residual Std. Error	2.077 (df = 483090)
F Statistic	3,593.707*** (df = 385; 483090)

Note: Data are from *Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014), Sarkees, Meredith Reid and Frank Wayman. (2010), and World Bank. (2023). Standard errors are reported in the parentheses.* *p **p ***p<0.01

Appendix 7

Extended regression on period 2007-2020.

Common Language and Bilateral Trade: Period 2007-2020

	Dependent variable:	
	Bilateral trade (log)	
	(1)	(2)
Common official language	0.486***	0.499***
	(0.019)	(0.019)
Common spoken language	0.408***	0.365***
	(0.033)	(0.034)
Common native language	0.432***	0.497***
	(0.046)	(0.047)
Linguistic proximity (1)	0.144***	
	(0.005)	
Linguistic proximity (2)		0.177***
		(0.007)
Distance (log)	-1.536***	-1.535***
	(0.006)	(0.006)
Common border	0.746***	0.763***
	(0.029)	(0.029)
Ex colonizer/colony	0.882***	0.844***
	(0.037)	(0.037)
Common colonizer	0.715***	0.700***
	(0.015)	(0.015)
Common religion	0.158***	0.201***
	(0.020)	(0.020)
Common legal system	-0.135***	-0.143***
	(0.010)	(0.010)
Years at war	-0.115***	-0.115***
	(0.017)	(0.017)
Observations	309,367	309,367
R^2	0.745	0.745
Adjusted R^2	0.745	0.745
Residual Std. Error (df = 308992)	2.103	2.104
F Statistic (df = 374; 308992)	2,418.324***	2,415.482***

Note: Data are from *Maddalena Conte, Pierre Cotterlaz & Thierry Mayer. (2022), Melitz, J. and Toubal, F. (2014) and Sarkees, Meredith Reid and Frank Wayman. (2010).*
Standard errors are reported in the parentheses. *p**p***p<0.01