Clustering by Linguistic Characteristics and International Trade

Alexandre Repkine\textsuperscript{a}

Associate Professor, Department of Economics, Konkuk University, Korea

Seung-Hoon Song\textsuperscript{b}

Associate Professor, Department of International Trade, Konkuk University, Korea

Contents

Abstract

I. Introduction

II. Why Does Linguistic Clustering Matter?

III. Data Sources and Summary

IV. Using Cluster Analysis to Group Countries According to Linguistic Dimensions

V. Empirical Methodology

VI. Empirical Results

VII. Conclusion

References

Abstract

This study focuses on the effects of similar language structures on bilateral trade between countries. While it has been established in the literature that countries sharing the same language trade more, we believe this is just a sub-case of a more general phenomenon that has to do with the fact that trade barriers are significantly reduced in case the negotiating parties share the same way of thinking. This in turn, we believe, is strongly linked to the language structure. By using a comprehensive database of language structures and bilateral trade flows, we confirm our hypothesis by finding that linguistic structural similarity is producing a strong effect on the international trade flows, which is independent of the extent to which the two trading partners' principal languages are mutually intelligible. This study emphasizes the need for more research linking the stock of knowledge accumulated by the science of psychology to economics, and the often overlooked importance of insights that can be suggested by cluster analysis.

Keywords: Cluster Analysis, International Trade, Language Structures

JEL Classifications: F14, O53, Z13

\* Received 17 November 2015, Revised 01 December 2015, Accepted 19 December 2015.
\textsuperscript{a} First Author, E-mail: repkine@konkuk.ac.kr
\textsuperscript{b} Corresponding Author, E-mail: steve13579@konkuk.ac.kr

© 2015 The Korea International Trade Research Institute. All rights reserved.
I. Introduction

In this study, we examine the effects of language clustering on bilateral international trade in the global context. While the effect of linguistic similarity between trading partners has been amply examined in the economics literature (see e.g. Melitz 2008), this similarity was usually understood in terms of the sameness of the languages spoken by the trading partners. Greenberg’s (1956) probabilities of two randomly chosen individuals in two trading countries speaking the same language became a popular way of measuring the extent of impediments to the international trade created by the linguistic differences. Higher values of such probabilities are normally taken to be indicators of lower transaction costs associated with the bilateral trade.

Clustering analysis is slowly making its way into the mainstream economic literature. Wooldridge (2003) convincingly describes how ignoring clusters may result in serious econometric problems such as biases in the OLS parameter estimates, or incorrect t-statistics. Moreover, taking an explicit account of clustering may provide economic insights that would be difficult to arrive at otherwise. Repkine (2012), for instance, examines how clustering of the East-Asian countries by the extent of similarity in their economic structures can be used for purposes of making economic policies directed at the creation of new free trading zones in the region.

The approach in Melitz (2008) where he uses a dummy to identify countries that widely use the same language such as Spain and Mexico, can be technically viewed as a clustering approach as well since his dummy indicates whether the two countries belong to the same cluster of countries speaking the same language. In this study, we extend this approach to define language clusters based on the similarity of their linguistic structures in general, which is not necessarily equivalent to the extent of their mutual intelligibility. The latter is a sufficient, but not necessary, condition for structural similarity. Two languages can be structurally similar such as e.g. Finnish and Hungarian, while at the same time being mutually totally unintelligible. We argue that structural similarity between the two principal languages widely spoken in the two trading partners will act as a factor decreasing transaction costs pertaining to bilateral trade, thus increasing its volume.

In Section II we discuss in detail why linguistic clustering would matter for the international trade. The general idea is, the notion of the geographical distance viewed in the gravity models framework as one major impediment to establishing trade links, can be extended to include linguistic differences that are not limited to the extent to which languages can be mutually intelligible. In fact, as we demonstrate in the empirical section of this study, mutual intelligibility produces a separate and independent effect on the international trade compared to similarity in the clustering sense. Since any language can be viewed as a multi-dimensional structure, clustering can be done along any combination of these dimensions, which is another issue we explore in this paper.

In the next section, we provide economic reasons as to why structural similarity between languages is important from the economic point of view. In Section III, we describe datasets that were employed in this study and provide summary statistics. Section IV explains how we assigned languages into
clusters. In Section V, we describe our empirical methodology, and Section VI contains empirical results. Section VII concludes.

II. Why Does Linguistic Clustering Matter?

We view linguistic similarity between two trading partners as one way in which the concept of geographical distance can be generalized, making it a perfect candidate for inclusion in the gravity models that explain international trade. The general idea is, international trade has many impediments to it, the geographical distance being just one more obvious of them. Linguistic differences create barriers to mutual understanding, successful negotiations, and maintaining business relationships, all of which work to increase the transaction costs. On the other hand, if the two trading parties speak the same language or are otherwise similar in the linguistic sense, international trade should be facilitated.

This last point is the main point of this study, namely, linguistic similarity may, but does not have to, have anything in common with the mutual intelligibility of languages spoken in the two trading countries. Melitz (2008) has comprehensively explored the many ways in which sharing a common language facilitates bilateral trade. He introduced the concept of “open circuit communication” defined as a situation where the same language is either official or widely spoken in the two trading partners’ countries, such as French between France and Belgium. The “direct communication” measure is defined as a probability that any two randomly chosen individuals in the two trading countries will speak the same language in case this language is spoken by at least 4% of the population in both countries. The probability in this case is computed as a product of the two population shares. Melitz (2008) finds that both direct and open circuit communication facilitate international trade, with direct communication producing the strongest effect.

The innovation of this study is to point out the fact that linguistic similarity may be present even in cases when mutual intelligibility is absent. Melitz’s (2008) direct and open circuit communication measures are just two ways to express the same idea, namely, that more mutual intelligibility facilitates trade. In this study, we posit that international trade is facilitated if the two trading parties’ way of thinking is similar, which does not necessarily involve the ability to understand each other without the help of an interpreter. Chen (2013) vividly demonstrates the importance of the way of thinking and its link to the language structure by examining the effects produced on economic behavior by the fact that in some languages, the present tense is associated with the future tense as well. He finds that the speakers of those “future-oriented” languages are also behaving in a future-oriented fashion, i.e. they save more, they retire wealthier, they take better care of their health and so on. We take on this idea to define similarity between a pair of languages in general as a Euclidean distance between the two vectors representing those languages’ structural characteristics. The idea is, the more structural difference there is between any two languages, the more different a the ways in which the trading
parties think, and therefore, the higher the transaction costs relative to the bilateral trade between them. In Section IV, we describe the way in which we form language clusters based on their structural similarity, and arrive at surprising associations such as the one between English and Chinese, that most people would not think of as being similar.

The main innovation of this study is to extend the idea of linguistic similarity beyond mutual intelligibility by embracing the structural similarity as well, which may or may not be associated with the ability of the two language native speakers to easily understand each other. Since there are many dimensions to any linguistic structure, we believe it is important to investigate in the future whether linguistic similarity along some particular dimensions is more important to the economic decision making compared to similarity along other dimensions. While we do not identify this sort of differences among linguistic dimensions in this paper, it is our strong belief that they will show up in other contexts.

III. Data Sources and Summary

We owe Frankel and Rose (2002) the dataset he made public for the replication purposes. This dataset contains observations on bilateral trade flows for the period between 1970 and 1995 for 185 countries. The observations are available at five-year intervals. Due to the unbalanced nature of the panel, some observations had to be deleted, leaving us with a total of 26501 observations.<Table 1> below provides summary statistics for the non-linguistic variables used in our analysis.

The variable Ln \( T_{ij} \) measures the volume of bilateral trade between countries \( i \) and \( j \), where \( T_{ij} \) is measured in millions of USD, Ln \( (RGDP_i \times RGDP_j) \) is the natural logarithm of the two countries’ real GDPs in billions of USD, and \( S_{ij} \) is the geographical distance between the two economies in kilometers.

The rest of the non-linguistic variables are dummies reflecting whether the two trading partners share a common language, have the same border, are members of the same currency union and/or a regional trade agreement, have had a common colonizer, have had a colonial past, and whether one of the two countries, or both of them, are an island nation. In the latter case, the ‘island’ variable is a category variable assuming the value of zero if none of the trading partners is an island nation, the value of one in case one trading economy is an island, and two if both of them are island economies.

We identify each country’s main language and then cluster languages into groups of linguistic similarity by employing two open-access databases, namely, World Atlas of Language Structures (WALS) and Ethnologue described below.

### Table 1. Data Summary for Non-linguistic Variables

| Variables | Mean | Standard Deviation | Minimum | Maximum |
|-----------|------|--------------------|---------|---------|
| Ln \( T_{ij} \) | 9.11 | 3.32 | 0.13 | 19.37 |
| Ln \( RGDP_i \times RGDP_j \) | 34.37 | 2.76 | 20.03 | 43.53 |
| Ln \( S_{ij} \) | 8.17 | 8.17 | 2.97 | 9.42 |
1. WALS Database

In order to assign the world’s countries into clusters according to the linguistic characteristics of the most widely spoken language, we used the World Atlas of Language Structures (WALS), the most comprehensive database to our knowledge containing information on 192 characteristics of 2680 languages. Dryer and Haspelmath (2013) contains a detailed description of the database that is a result of collaboration among many scholars. The linguistic characteristics are structured in twelve broadly defined categories such as e. g. phonology, morphology and nominal syntax.

2. Ethnologue

Since our linguistic clustering procedure is based on the comparison of language characteristics between two languages in two different countries, we have to deal with a problem posed by the fact that in many countries more than one language is spoken. In Belgium, for instance, they speak Flemish and Dutch, while in South Africa the number of spoken languages exceeds one hundred. We dealt with this problem by assigning to each country a principal language defined so by the Ethnologue database in Lewis, Simons and Fennig (Eds.) (2015). This database is now being widely used by economists and linguists alike, providing among other things the number of speakers for each language spoken in the world’s countries. Ethnologue provides for each country the name of the principal language. If two or more languages are listed as principal, e.g. Belgium or Singapore, we selected the one with the largest number of speakers. Very often, in the case of the African countries such as Cameroon or Angola, the number of speakers of indigenous languages is small compared to the total population so none of the originally spoken languages can be said to be dominating the linguistic landscape. In this case, we look at the national language (typically one of those brought by the former colonizers such as English, French or Portuguese) with the largest number of speakers.

IV. Using Cluster Analysis to Group Countries According to Linguistic Dimensions

Cluster analysis is a technique of dividing multi-attribute objects into groups according to some measure of similarity between those objects, see e. g. Anderberg (1973). Repkine (2012), for instance, employed this approach in order to determine clusters of East Asian economies needed in their analysis of the trade blocs formation. We group the world’s languages into clusters by using the WALS database that assigns 192 attributes to each language, using as a distance between any couple of languages the Euclidean norm of the 192-dimensional vector of differences between the values of the corresponding attributes.

There exist two major approaches to clustering multi-attribute objects based on how we determine the number of groups into which we cluster our observations, see e. g. Everitt, Landau and Leese (2001). The first approach is a hierarchical one in the sense that it starts by considering each observation as a cluster of its own, gradually adding the most similar observations to form larger-sized
groups until the whole sample becomes one big cluster. The non-hierarchical methods distribute observations into a predefined number of groups, e.g. fifteen. The clusters are then formed by an algorithm that "seeks" to minimize the representative distance of the individual observations in a cluster from its centroid. The representative distance is most often understood in terms of either a mean or a median distance. Since there is little theoretical guidance as to the number of clusters into which languages should be placed, we experimented by starting with a small number of clusters (five) and kept on increasing the number of clusters until it became sufficiently large, stopping at forty.

We then applied a series of statistical tests to converge on the optimal number of clusters. These tests are based on the hierarchical approach to clustering that employs certain stopping rules that are used to stop the agglomeration process. In this study, we employ the two more popular rules, namely, pseudo-F index (Calinski and Harabasz, 1974), and the Duda-Hart index (see Duda, Hart and Stork, 2001). Both kinds of tests produce values of their corresponding statistics for each number of clusters with the highest values of these statistics indicating the most distinct clustering. <Table 2> below represents the results of this exercise for the five alternative ways in which similarity between observations can be defined. Thus, the single-linkage method defines similarity in terms of the shortest distance between the two observations in two distinct clusters. The complete linkage approach looks at the minimum diameter of the sphere that enshrouds all observations in a pair of clusters. The average linkage, centroid and Ward’s similarity is defined in terms of the average distance between the clusters, the distance between clusters’ centroids, and the sum of squared distances within the clusters taken over all variables. More details on the different types of linkages and similarities can be found in Everitt et al. (2001).

We observe that in the overwhelming
majority of the cases, both pseudo-F and Duda-Hart tests determine forty to be the optimal number of clusters in our language sample. Thirty-nine clusters are chosen by these two tests when average linkage is used as a measure of similarity, and in two cases, the optimal number of clusters is chosen to be thirty-six and thirty-seven. We therefore decide to conduct our further analysis dividing languages into forty clusters.

Now that we have determined the optimal number of clusters, our next step is to assign each language to a specific cluster with the view of creating a dummy identifying the two trading partners whose main languages belong to the same cluster. We assign languages to clusters by applying the two widely used methodologies, namely, the k-means and k-medians approach. In both cases, the number of clusters has to be known in advance, which we determined to be forty as discussed above. The idea behind both methods is to minimize the overall distance within each cluster of the individual observations from the cluster’s centroid. The overall distance is understood in terms of the means and medians of the individual distances for the k-means and k-medians methodologies, respectively. <Table 3> below presents some characteristic languages and their respective clusters:

We see from <Table 3> above that language classifications suggested by both k-means and k-medians methodologies are rather similar for the reported languages, which is a general characteristic of the sample. Expectedly, the Slavic languages are grouped together, although the Polish language forms a cluster of its own according to the k-medians approach. Similarly, it is no surprise to find Finnish and Hungarian in the same cluster in both cases since these languages’ origin in the Finno-Ugric family is well-known. What is remarkable is that these two languages are mutually and totally unintelligible. However, what matters in the context of this study is the way of thinking underlying these two languages, which is very similar. Korean and Japanese provide us with another example of two mutually unintelligible languages that should be placed in the same group since their grammatical similarity as well as heavy reliance on the Chinese lexicon is very well known. It is remarkable that linguists are still arguing whether Korean and Japanese should be included in the same language family or not. The African languages are all grouped together in Cluster 9 according to the k-means

| K-means Cluster Number | Clustering by K-means | K-medians Cluster Number | Clustering by K-medians |
|------------------------|-----------------------|--------------------------|-------------------------|
| 13                     | Czech, Polish, Serbian| 20                       | Serbian, Czech          |
| 14                     | Finnish, Hungarian    | 26                       | Finnish, Hungarian      |
| 16                     | Korean, Japanese      | 11                       | Korean, Japanese        |
| 7                      | Swedish, Danish, Icelandic, Norwegian | 16 | Danish, Swedish |
| 9                      | Mossi, Dzonkha, Fulfule, Bemba, Rundi, Dhivehi | 35 | Norwegian, Romanian, Pashto |
| 6                      | English, Mandarin     | 2                        | English, Mandarin       |

Source: Author’s calculations.
approach, which is again not surprising.

What is really interesting is that English and Mandarin Chinese are consistently grouped in one two-member cluster by both methodologies! Different as these two languages are in pronunciation and word origin, many learners of both languages noticed a similarity in the two languages' grammar, the latter being highly analytical in nature. Some clusters are surprising e.g., Cluster 35 obtained by the k-medians approach where Norwegian is grouped together with Romanian and Pashto. While we have no explanation for this grouping at the moment, we believe it is an interesting issue to be explored by the linguists.

V. Empirical Methodology

In order to gain inference into the effects of a common language cluster on bilateral international trade, we make use of the well-established gravity model. Although the theoretical foundations for this model are still a subject of debate, it has been long known and mentioned by Anderson (1979) that international trade flows are explained surprisingly well by it. The general idea is rather intuitive, i.e., similar countries trade more. Following Melitz (2008), we control for the fact that the two trading partners might share a common language such as Belgium and France where in the former country, French is one of the official languages.

However, in this study, we broaden the idea of linguistic similarity to account for the fact that languages belonging to the same linguistic cluster reflect similarity in terms of way of thinking, which may be an important determinant of trade. As mentioned in the previous section, belonging to the same cluster does not necessarily mean that the two languages are “officially” recognized by the linguists to be similar or to even belong to the same language family. As demonstrated in Table 3, English and Chinese are consistently clustered together, while in addition to being mutually totally unintelligible, the two languages do not even belong to a common macrofamily such as Indo-European. For that reason, we expect language similarity and language clustering to exhibit two independent effects.

Apart from controlling for the size of the two trading partners, geographical distance between and adjacency of the two trading partners, we include additional controls inspired by Melitz (2008) such as belonging to a currency or political union, free trade area, ex-colonial past, and ex-common colonizer.

Helpman, Melitz and Rubinstein (2006) argue in favor of gravity equations as a preferred analytical tool for gaining inference on the determinants of the international trade flows. In particular, their study emphasizes the importance of including fixed effects into the empirical specification since ignoring these effects is tantamount to assuming homogeneity among country pairs with respect to firms’ productivity and the fixed costs of firms’ exporting. We therefore estimate the following specification of the gravity model:

\[
\ln(T_{ij}) = \alpha_0 + \alpha_1 \ln(RGDPS_i \times RGDPS_j) + \alpha_2 \ln(S_{ij}) + \alpha_3 D_{ij} + \alpha_4 D_{ij} + \epsilon_{ij} \tag{1}
\]

where \(i\) and \(j\) are index countries, \(RGDP_i\) stands for real GDP of country \(i\), \(S_{ij}\) is the geographical distance between the two
countries’ capitals, $\overline{D}_{ij}$ is the vector of non-linguistic dummies such as common ex-colonizer, and $L_{ij}$ is the vector of the linguistic variable such as common language and common language cluster dummy. The $t$ index refers to time.

As demonstrated in the previous section, we empirically estimate the optimal number of language clusters for our sample to be forty. We employ the non-hierarchical grouping methodologies, namely, $k$-means and $k$-medians, to group countries into clusters according to the main language spoken in those countries. To check for the robustness of our findings, we use distance measures other than the Euclidean one that are often used in cluster analysis such as Manhattan or Canberra measures.

**VI. Empirical Results**

**1. Gravity Model Regression**

We start by running a gravity model regression in (1) focusing on the effects of the two linguistic variables in our study, namely, the common language and common cluster. The main idea is to see whether, indeed, belonging to the same cluster means something quite different from using a common language. A separate statistically significant effect of linguistic clustering would corroborate our hypothesis that languages that are similar in a structural fashion account for similar ways of thinking in the two trading partners even if the principal languages spoken in these trading economies are mutually unintelligible and/or belong to two different language families. In order to single out the effect produced by the linguistic clustering we applied two clustering procedures to our dataset that resulted in a dummy for a pair of languages signifying their belonging to the same linguistic cluster. The two more popular procedures to produce such clustering are the $k$-means and $k$-medians methodologies. Both are based on the idea of minimizing an average distance between observations within the same cluster. It is the definition of the “average” that results in alternative clustering, i.e. the $k$-means procedure minimizes the mean distance, while the $k$-medians procedure minimizes the median one. In case sample languages form a continuum, i.e. if for each language in the sample it is possible to indicate a closely related and mutually intelligible language, the two procedures result in almost identical clustering outcome. However, since this is not the case for the real world languages, including those in our sample, the two clustering procedures will in general produce different results. By studying the clustering effects using two alternative definitions of clustering, we are checking for the extent to which our findings are robust to the clustering procedure specification. Table 4 > below presents our empirical results that were obtained using the $k$-medians approach to the clustering of languages.

We observe that, in line with what one would generally expect from estimating a gravity model, all parameters have expected signs. Thus, trade is benefited by larger economic sizes of the trading partners and by geographical proximity. Belonging to the same currency union or a regional trade agreement boosts international trade as well. Having a common border does not appear to be important in all specifications but the last one where the corresponding parameter becomes
importantly to two partners related initial hypothesis. Finally, does not produce a positive effect on bilateral trade, while the presence of a common colonizer does not appear to be an important factor. Finally, being an island country results in increased trade volumes.

Our empirical results appear to support the initial hypothesis of the two separate effects related to language similarity: two trading partners sharing a common language, and the two economies’ principal languages belonging to the same cluster. The latter case is importantly distinct from the former, which is greatly illustrated by the U.S. and China. As already mentioned previously, the two countries do not share a common language at all, however, according to our clustering analysis, English and Mandarin Chinese are consistently placed into the same group by our clustering procedures. The message is, even if English and Chinese are mutually not intelligible, the similarity of the two languages’ structures provide for a greater international trade potential due to the similarity in the way of thinking.

We observe that, while the effect of a common language is estimated to be positive and statistically significant in all specifications,
Clustering by Linguistic Characteristics and International Trade

Table 5. Bilateral Trade and Language Clustering: K-Means Method

|               | (1)     | (2)     | (3)     | (4)     | (5)     | (6)     | (7)     |
|---------------|---------|---------|---------|---------|---------|---------|---------|
| Constant      | -9.061*** | -10.35*** | -10.42*** | -10.44*** | -10.92*** | -10.64*** | -12.162*** |
|               | (0.208)  | (0.215)  | (0.216)  | (0.219)  | (0.224)  | (0.224)  | (0.235)  |
| ln RGDP \times | 0.842*** | 0.860*** | 0.860*** | 0.860*** | 0.867*** | 0.860*** | 0.907*** |
| RGDP         | (0.005)  | (0.005)  | (0.005)  | (0.005)  | (0.005)  | (0.005)  | (0.006)  |
| ln S_{ij}    | -1.278*** | -1.214*** | -1.206*** | -1.203*** | -1.176*** | -1.180*** | -1.223*** |
|               | (0.017)  | (0.018)  | (0.018)  | (0.019)  | (0.019)  | (0.019)  | (0.019)  |
| Common Border | 0.031    | 0.054    | 0.111    | 0.249*** |
|              | (0.096)  | (0.095)  | (0.095)  | (0.094)  |
| Currency Union| 1.419*** | 1.443*** | 1.364*** |
|              | (0.149)  | (0.148)  | (0.148)  |
| Colonial Past | 2.277*** | 1.995*** |
|              | (0.123)  | (0.123)  |
| Common Colonizer| 0.051    |
| Island Category| 0.592*** |
| Regional Trade Agreement | 1.483*** |
|               | (0.101)  |

Note: standard errors in parentheses, *** stands for a 1% significance level, ** for 5%, and * for 10%. Number of observations: 26501.

its magnitude drops significantly by almost fifty percent as we add more regressors to the basic specification (1). The same is happening to the coefficient on language clustering which sheds about a quarter of its value in specification (7) compared to (3). We find it remarkable that there is nosingle specification where the language clustering dummy would come out insignificant or with a negative sign, which we believe is a strong evidence in favor of the two linguistic effects co-existing with each other and working to increase the volume of bilateral trade through two independent channels.

2. Robustness Checks

In order to see whether our empirical results are robust in terms of the choice of clustering procedure, we repeat our analysis by applying the k-means, rather than k-medians, procedure. That is, rather than forming clusters on the basis of the minimum average median distance within each clusters between individual observations and the duster centroid, we do the same for the means. Table 5 below presents the results. The major result we are finding is that linguistic clustering is producing a statistically significant positive effect on the volume of
Table 6. Bilateral Trade and Language Clustering by Alternative Within-Cluster Distances

| Distances                | Manhattan | Maximum Value | Canberra |
|--------------------------|-----------|---------------|----------|
|                          | k-median  | k-means       | k-median | k-means |
| Constant                 | -12.681***| -12.703***| -12.650***| -12.686***| -12.655***| -12.655***|
| Ln RGDP × RGDP<sub>i</sub> | 0.904***  | 0.904***| 0.904***| 0.904***| 0.904***| 0.904***|
| Ln S<sub>j</sub>         | -1.147*** | -1.145***| -1.150***| -1.147***| -1.149***| -1.149***|
| Common Border            | 0.246***  | 0.247***| 0.247***| 0.243***| 0.250***| 0.252**|
| Currency Union           | 1.120***  | 1.120***| 1.121***| 1.118***| 1.122***| 1.126***|
| Colonial Past            | 2.067***  | 2.066***| 2.069***| 2.065***| 2.069***| 2.069***|
| Common                   | 0.054***  | 0.059***| 0.053***| 0.058***| 0.051***| 0.051***|
| Colonizer                | 0.537***  | 0.538***| 0.536***| 0.538***| 0.536***| 0.537***|
| Island Category          | 1.490***  | 1.477***| 1.485***| 1.486***| 1.488***| 1.483***|
| Regional Trade Agreement | 0.465***  | 0.456***| 0.484***| 0.461***| 0.482***| 0.423***|
| Same Cluster             | 0.212***  | 0.232***| 0.176***| 0.220***| 0.178***| 0.160***|
| R-squared                | 0.534      | 0.545      | 0.544      | 0.545      | 0.544      | 0.544      |
| Root MSE                 | 2.227      | 2.227      | 2.227      | 2.227      | 2.227      | 0.223      |

Note: standard errors in parentheses, *** stands for a 1% significance level, ** for 5%, and * for 10%. Number of observations: 26501.

international trade, which is separate from the common language effect.

We do not observe any major difference in the sign or magnitude of the estimated parameters when we apply the k-means as opposed to the k-medians procedure in clustering languages. Most importantly, the effects of both the common language and the same cluster remain positive and statistically significant, corroborating our hypothesis that these are two distinct effects.

Another way to check for the robustness of our findings is to look at the effects produced by the way one defines a distance between two observations. As mentioned already, we decided to use the Euclidean distance. However, several alternative distances are available. For instance, it is possible to apply the absolute value distance, otherwise known as Manhattan distance. Another alternative is to use the maximum value distance, in which case, clusters are formed by minimizing the maximum distance within each cluster between its centroid and the most “remote” observation. Another popular option is to use the Canberra distance which is a weighted version of the Manhattan distance where the weights are computed as sums of the absolute values of two corresponding components in the compared vectors. <Table 6> presents the results of estimating specification (7) in <Tables 4> and <Table 5>, which is the most comprehensive one.
As demonstrated by the results in <Table 6>, the choice of a particular type of distance to be used by a clustering procedure does not produce any sizable influence on the sign or significance of the estimated parameters.

3. Relative Importance of Certain Language Characteristics

As we mentioned already, the WALS database employed in this study contains data on 192 linguistic characteristics on each one of its languages. These characteristics are grouped into twelve different categories such as morphology or nominal syntax. While our empirical results suggest that if the two trading partners’ principal languages belong to the same cluster and their bilateral trade volume increases, it is natural to ask whether clustering by a specific linguistic category is of utmost importance to international trade.

The language structure categories referred to in <Table 7> are taken from the original WALS database. Morphology classifies languages by the extent to which they are inflective, e.g. whether they have cases or whether their verbs are inflected. Nominal categories refer to the number of noun genders, nominal plurality and the general extent to which nouns are modified. Adjectives and verbal categories refer to the way in
which adjectives, for instance, may act as verbs, and to the number of verb tenses. Word order and simple clauses categorize the organization of simple sentences with respect to the word order, i.e. subject-object-verb versus subject-verb-object. Finally, complex sentences and lexicon capture structural similarity by comparing the organization of complex sentences and the lexical differences between, say, hand and arm.

While the structural categories we chose are rather arbitrary, the point of this exercise was to examine whether such a choice affects our main results. We experimented with several other groupings of the linguistic characteristics with no implications for our main results. However, we believe it is essential to investigate in the future the economic effects of grouping languages along various dimensions.

Due to the problem of frequent ties associated with the insufficient number of observations on specific categories we had to limit our analysis to the case of four language characteristics rather than the original twelve using the k-medians approach to the calculation of the overall distance. <Table 7> below presents the results.

The most noticeable result implied by <Table 7> is that regardless of the type of a language characteristic, the size of the effect of both common language and the same cluster on bilateral trade is virtually the same. Moreover, the magnitudes of these two effects are estimated to be very close to their counterparts estimated on the basis of clustering by all language characteristics together across various specifications. This implies that the effects produced by the language structure similarity on bilateral trade volume are reflecting a similarity between languages that is distributed across different linguistic categories in rather uniform a fashion. For instance, if the two languages are structurally similar in terms of their morphology and syntax, they will also be similar in terms of the structure of their simple clauses and complex sentences. The source of such uniformity would be interesting to establish, but it is outside of the scope of the present analysis.

VII. Conclusion

In this study, we extend the idea of geographical distance between trading partners to accommodate the concept of structural similarity between different languages. While our empirical results are in line with the well-known findings of Melitz (2008) on the significant effect exerted by the fact that two trading partners widely speak the same language, we identify a separate effect of the two countries’ languages belonging to the same cluster. The latter is defined in terms of the structural similarity that we infer on the basis of almost two hundred characteristics pertaining to each language in the WALS database by Dryer and Haspelmath (2013).

A striking finding of our research is that languages normally thought of as being entirely different such as Chinese and English in fact share a lot of structural similarities, which is reflected in the fact that alternative clustering procedures place those two languages in the same cluster. We argue this sort of similarity is accounting for increased volumes of bilateral trade since similar language structures act as factors decreasing the negotiations, and transaction costs in
general of international trade. The idiosyncratic way of thinking characterizing the negotiating parties appears in many ways to be caused by the language structure. The innovation of this study is to point out the fact that similar language structures do not necessarily imply mutual intelligibility, which appears to be the main assumption underlying the extant research linking languages and international trade.

We find that structural similarity in the linguistic sense produces a positive effect on international trade volumes for the sample of almost two hundred countries covering the period from 1970 to 1995. One of our findings is that the effect of this structural similarity is fairly uniform in the sense that it does not appear to depend on a specific category by which similarity may be judged such as syntax, morphology, or whether the words for “hand” and “arm” are different in certain language pairs. This finding is robust with respect to the definition of distance between languages and the choice of the clustering rules.

We conclude therefore that sharing a common language and having a principal language that belongs to the same language cluster with the principal language cluster of one’s trading partner are two separate effects that are both producing a positive effect on the volume of bilateral trade. We also observe that the relative size of these two effects remains relatively constant over various specifications. Thus, in the most comprehensive specification sharing a common language produce twice as strong an effect compared to the effect produced by a common language cluster.

More work needs to be done in the area of linking the stock of knowledge in psychology and economics since it is far from dear yet what it is exactly that makes trading partners who share structurally similar languages to enjoy lower transaction costs. The role of cluster analysis in economics should gain more prominence as more empirical evidence suggesting its importance is entering the mainstream of economics science. We believe this study is one of the first steps in this direction.

References

Anderberg, M. R. (1973), *Cluster Analysis for Applications*, New York, NY: Academic Press.

Anderson, J. (1979), “A Theoretical Foundation for the Gravity Equation”, *American Economic Review*, 69, 106-116.

Calinski, T. and J. Harabasz (1974), “A Dendrite Method for Cluster Analysis”, *Communications in Statistics*, 3, 1-27.

Chen, M. K. (2013), “The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets”, *American Economic Review*, 103(2), 690-731.

Dryer, M. S. and M. Haspelmath (Eds.) (2013), *The World Atlas of Language Structures Online*, Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. Available from http://wals.info (accessed October 25, 2015).

Duda, R. O., P. E. Hart and D. G. Stork
(2001), *Pattern Classification and Scene Analysis* (2nd ed.), New York, NY: Wiley.

Everitt, B. S., S. Landau and M. Leese (2001), *Cluster Analysis* (4th ed.), London: Arnold.

Frankel, J. and A. Rose (2002), “An Estimate of the Effect of Common Currencies on Trade and Income”, *Quarterly Journal of Economics*, 117, 437-466.

Greenberg, J. H. (1956), “The Measurement of Linguistic Diversity”, *Language* 32(1), 109-115.

Helpman, E., M Melitz and Y. Rubinstein (2008), “Trading Partners and Trading Volumes”, *Quarterly Journal of Economics*, 123(2), 441-487.

Lewis, M. P., G. F. Simons and C. D. Fennig (Eds.) (2015), *Ethnologue: Languages of the World* (18th ed.), Dallas, TX: SIL International. Available from http://www.ethnologue.com (accessed October 25, 2015).

Melitz, J. (2008), “Language and Foreign Trade”, *European Economic Review*, 52, 667-699.

Repkine, A. (2012), “How Similar Are the East Asian Economies? A Cluster Analysis Perspective on Economic Cooperation in the Region”, *Journal of International and Area Studies*, 19(1), 27-44.

Wooldridge, J. M. (2003), “Cluster-sample Methods in Applied Econometrics”, *American Economic Review*, 93(2), 133-138.