Trading with "Our Own People":

Linguistic Similarity, Social Identity and International Trade

A thesis submitted by

Xiaoyang (Derrick) Ji

in partial fulfillment of the requirements for the degree of

Master of Science

in

Economics

Tufts University

Graduate School of Arts and Sciences

May 2021

Advisor: Enrico Spolaore

Abstract

This paper investigates how common dialects spoken in China and Southeast Asian countries and the social group identity shaped by these dialects affect trade patterns. Using a gravity model, I construct a dialect similarity index and estimate its influence on exports from Chinese provinces to Southeast Asian countries. The results show that linguistic similarity has a positive and significant effect on trade. I also explore several mechanisms behind this fact, including one based on the social identity theory.

Acknowledgment

I am very grateful to Enrico Spolaore and Federico Esposito for their guidance and support. I also wish to thank Melissa McInerney, Marcelo Bianconi and other seminar participants for their helpful suggestions and comments. All errors are my own.

1. Introduction

"Ga di nang" is a term Teochew people[1] often use to refer to fellow Teochew people, meaning "our own people" and taking it to a deeper level – "brotherhood". Such a sense of closeness can often be felt when people meet others who also speak the same language, no matter which language they speak. Will such kind of bond between people who speak the same language influence trade? If so, are there further reasons beyond language similarity?

Language plays an important role in people's self-identification of ethnic groups. People who speak similar languages tend to consider themselves in the same ethnic group. Fearon (2003) showed that relationships between languages can be used as a proxy for cultural similarity and thus can be used in ethnic group classification. Mai (1985) also showed that people in Southeast Asian countries such as Singapore and Malaysia classify themselves into different ethnic groups based on what dialects they spoke.

In this paper, I study how language-based ethnic group identity influences trade patterns between each province of China and each Southeast Asian country (province-country pair). When it comes to the language spoken in China, many people think it is Mandarin. However, Mandarin is just the official language of China, and there are also many dialects spoken in different regions of China. According to Xu and Miyata (1999), there are 17 Chinese dialects (such as Cantonese and Hakka) and 105 sub-dialects (such as Hokkien and Teochew). These dialects and sub-dialects are in fact mutually unintelligible and hence can be treated as different languages.

What's more, some of these Chinese dialects are also spoken in Southeast Asian countries. This is because many people in Southeast Asian countries are immigrants

---

[1] People who speak Teochew Chinese dialect.

from South China, which can be dated back to the 15th century, according to Li (2012). As a result, many of them speak different Chinese dialects and classify themselves into different ethnic groups. Take Singapore as an example. According to Singapore Census of Population (2010), the population in Singapore is about 4.59 million, and among them about 29.66% are from the ethnic group speaking Hokkien, 14.90% Teochew, 10.83% Cantonese, 6.18% Hakka and 4.71% Hainanese.

In addition, these Southeast Asian countries have a close trade relationship with China; the Association of Southeast Asian Nations (ASEAN) already becomes China's largest trading partner. In 2019, the trade volume between China and ASEAN exceeded $641 billion, which is more than that between China and the US. The strong trade ties between those Southeast Asian countries and different provinces of China, together with the fact that several common dialects spoken in different countries and Chinese provinces, make this area a good laboratory to study the effect of language-based ethnic group identity on trade and the specific mechanisms behind it.

This paper makes several contributions to the literature. First, compared to the conventionally used binary variable indicating common language, the index I construct in this paper can measure language similarity more precisely. Second, compared to other literature on language and trade, I investigate this issue from a broader angle of social group identification, and thereby explore the mechanisms more thoroughly. In particular, I show a new mechanism which is based on the social identity theory.

The remainder of the paper is structured as follows. In Section 2 I briefly describe the data. In Section 3 I estimate the effect of language similarity on trade. In Section 4, I discuss several possible mechanisms about the estimation result. And Section 5 concludes.

2. Data

In this paper, I use province-country pair trade data between 31 provinces of Mainland China and 8 Southeast Asian countries. The list of the 31 China provinces and 8 Southeast Asian countries is provided in Table A1 and A2, respectively. For the trade data, I use both the export and import flows from a Chinese province to a Southeast Asian country. I then aggregate the trade data, language data as well as other supplementary data (such as GDP, population and distance data) into a short panel for 12 quarters (2017Q1 – 2019Q4).

For the language data, I focus on the following 5 Chinese dialects: Cantonese, Teochew, Hakka, Hokkien and Hainanese, which are the main dialects spoken in Southeast China and other Southeast-Asian countries. I assume that people are indifferent from dialects other than their own, so other dialects can be treated as a homogenous "Other Dialect" from the perspective of those who do not understand them. And because the available language data is less frequent than other economic indicators like GDP and exports, I make another assumption that the percentage of people speaking a certain dialect in a certain region remains unchanged in a relatively short period of time. This assumption is reasonable if (1) the growth rate of people speaking different dialects/languages are similar, and (2) there are not many people migrating across different regions, or the migration is not biased toward people speaking certain dialects/languages. I assume condition (1) holds, at least in the recent decades. Condition (2) also holds since in recent decades there is no massive emigration from China to Southeast Asian countries[2], there are no records of massive emigration among those Southeast Asian countries, and although there is some inter-province migration in

[2] The majority of emigration from China to Southeast Asian countries happened between early 17th century and 1950s, and the emigration flows afterwards are much smaller, according to Zhuang (2008).

China, these migration flows do not have large effects on the distribution of dialects in China[3]. As a result, to overcome the data availability problem, I use recent dialects data as an estimation of the dialects data in the three years (2017-2019) of interest.

To measure the common language-based ethnic group identity within each province-country pair, I construct a language similarity index, which is defined as the following:

$$\text{languageSimilarity}_{\text{ij}} = \sum_d popShare_i^d * popShare_j^d,$$

where $popShare_k^d$ is the percentage of people speaking a certain dialect d in economy k, and $d \in \{\text{Cantonese}, \text{Teochew}, \text{Hakka}, \text{Hokkien}, \text{Hainanese}\}$. This language similarity index represents the joint probability of if you randomly pick a person from a province i in China, and randomly pick another person from a certain Southeast Asian country j, and both of them speak the same dialect.

Other data used in estimation include GDP, population, and geographic data such as distance, landlock and common border. See Table A3 for description of main variables, Table A4 for summary statistics and Table A5 for correlation between language similarity index and other control variables. And for more details about the data sources, see Appendix A.

3. Estimation

To estimate the effect of language-based ethnic group identity on trade, I use the gravity equation which states that the volume of trade between two economies is increasing in their economic sizes and decreasing in the geographic distance between

---

[3] In fact, according to Li et al. (2015), most of the migration flows are from middle/western provinces to eastern provinces, which would lower the population share of those dialects spoken in Southeast China, and hence cause the underestimation of the effect of language similarity on trade.

them. The gravity equation was first introduced by Tinbergen (1962), and performed well in international trade empirical studies. Then there are researches aiming to provide a theoretic foundation for the gravity equation. For instance, Eaton and Kortum (2002) developed a Ricardian trade model that incorporates geographic distance into general equilibrium. Anderson and van Wincoop (2003) highlighted the importance of adding multilateral resistance terms into a gravity model. Chaney (2008) incorporated heterogeneous firms into the determination of bilateral trade flows.

In the remaining part of this section, I firstly do the estimation using a naïve gravity equation, and then a structural one. In the end, I also do some robustness checks.

3.1 Naive gravity equation

$$lnX_{ijt} = \beta_0 + \beta_1 lnGDP_{it} + \beta_2 lnGDP_{jt} + \beta_3 lnPop_{it} + \beta_4 lnPop_{jt} + \beta_5 lnDist_{ij} + \beta_6 commonBorder_{ij} + \beta_7 landlock_{ij} + \beta_8 languageSimilarity_{ij} + \beta_9 I_t + u_{ijt} \quad (1)$$

For the naive form of gravity equation, I use the following equation[4]:

I firstly do a log-linear regression, and the result is shown in Table 1. From the result we can see that when only using the GDP of both economies and distance and controlling for common border and landlock, the distance elasticity of trade is -0.934, which is close to -1, consistent with the results of many other empirical studies on the gravity equation. However, when further controlling for the population of both economies, the distance elasticity is reduced drastically to -0.647. And when finally adding the language similarity index, the distance elasticity is further reduced to -0.596, which is only roughly 60% of the original value. Note that these additional controls are all related to social interactions between two economies; holding all else equal, larger

[4] There is no need to control for tariff in this equation, because the ASEAN–China Free Trade Area (ACFTA) has been established since 2010, and the tariffs on the majority of the goods traded among China and ASEAN countries have been eliminated to 0.

population and more similar language both can lead to more social connections between the two economies. This result indicates that with the interaction of the people between the two economies, the distance barrier becomes less important. And from column 3a, we can see that the beta coefficient of language similarity is 0.073, while that of distance is -0.087, so the impact of language similarity on trade is roughly 85% of that of distance.

But there exists a “zero trade” issue in the sample, and a log-linear regression will omit observations in which trade flows are zero, which will cause selection bias if these zeros are not random missing data or random rounding errors. Note, however, that even if the selection bias exists, it is not severe in the dataset here, because China has a strong tie with Southeast Asian countries in trade, and in the 31(provinces) * 8(countries) * 12(quarters) * 2= 5952 observations, there are only 550 observations (79 in export flows and 471 in import flows) with zero trade, which accounts for only 550/5952 = 9.24% of the whole sample. But even so, I still seek to tackle with the “zero trade” issue. And to achieve this goal, we need to examine the causes of it first.

According to Head and Mayer (2014), there are two possible reasons for zero trade flows: (1) “statistical zeros”: the zeros occur due to missing data or rounding/ declaration thresholds; (2) “structural zeros”: although structural gravity models predict that trade flows are always strictly positive, in the real world, because of the existence of fixed export cost and the finite number of firms and consumers, it is possible that some economies do not trade with each other.

But for the data used in this paper, the “statistical zeros” case is unlikely. First, structural trade models imply that zero trade flows are more likely when bilateral trade is expected to be low, such as when the two economies are distant from each other and/or their sizes are small. To see more clearly which factors should account for the extensive margins of trade, I run the Probit regression based on equation (1)[5], and the result is recorded in Table 2 (Column 1 reports marginal effects, while Column 1a standardizes these marginal effects to report the contribution of one standard deviation increase in a certain variable to the increase in the probability of observing a positive trade flow.) The result shows that the GDP of the two economies have the largest significant effect on the extensive margin. I also have a close look at the those zero-trade observations. From Table 3, we can see that all 79 zero export flows occur between economies with at least one of the two having the smallest or second smallest GDP. Also note that for the province-country pair of "Tibet-Brunei", both of them having the smallest GDP among the provinces/countries in the dataset, contributes the most zero trade observations. And for the 471 zero import flows (not shown in the table), 96.60% of them also involve economies with the smallest or second smallest GDP. Hence, these zeros are unlikely to be random missing data. In addition, these zeros are not caused by rounding/ declaration thresholds, because the data are recorded in the unit of dollars. And we can see this fact from Table 4, which shows that there are 17 observations with trade flow less than 100 dollars and they all involve with at least one smallest or second smallest economy. As a result, we can safely conclude that these zeros are in fact "structural zeros" and omitting these zeros will cause selection bias (though it might not be large).

To deal with the "structural zeros", one commonly used method is to add one to observed exports before taking logs. One critique of this method is that estimation results depend on the unit of measurement. Head and Mayer (2014) reported that "distance elasticities range from -1.93 to -0.09 as we change the exports units from

[5] This is similar to the first stage of the two-state estimation approach proposed by Helpman, Melitz and Rubinstein (2008), but because the religion data are not available for all economies involved, I do not carry out their approach.

dollars to billions of dollars. The estimated impact of common currencies switches from negative and significant to positive and significant simply by changing units from millions to billions." However, this critique and those evidence are not well-grounded. With "structural zeros", we are assuming that the "expected flows" are low, hence when we add one to observed exports, we are hoping to bypass the undefined "ln(0)" but at the same time still keep the data as similar to the actual data as possible. As a result, adding one dollar is a reasonable choice, which leaves the majority of export values almost unchanged, but adding one million dollars to each export value is very likely to change the distribution of export values, let alone adding one billion dollars. What's more, from Table 4 we can see that the lowest export value is 1 dollar[6], therefore assuming the unobserved low export value as 1 dollar will not overestimate the actual value. In addition, from Table 3 we can see that even for the province-country pair of "Tibet-Brunei" which has the most zero trade observations, they still trade sometimes, which means that the fixed and variable costs for any province-country pair in the dataset are not prohibitive, and hence assuming the unobserved low export value as 1 (or 0) will not lead to large bias to the estimation of trade cost.

From the above analysis, we know that adding one dollar (while using a larger measurement unit is less justified) to observed exports before taking logs is a reasonable method to deal with zeros in trade data, especially for the dataset used in this paper. And I further test the influence of different measurement units, and from Table 5 we can see that changing units from dollar to thousands to millions only leads to a small variation in the estimated effect of language similarity, which proves that adding one before taking logs works well in the dataset here. This result also shows that a small

[6] This value is quite small but is still possible. Fernandes et al. (2019) also report that in a database featuring firm-level exports from 50 countries, they observed many firms with very small export sales (sometimes as low as $1).

proportion of zeros indeed causes little selection bias in the estimations. And even though different measurement units do not influence the estimation much, in the following analysis I still choose dollar as the unit so as to keep the distribution of export values as unchanged as possible.

To deal with the “structural zeros”, there are also other approaches. One approach is to use the Tobit estimator with left-censoring at zero on the log of trade plus a constant (I still use one.) Another approach is to use the Pseudo Poisson Maximum Likelihood estimator. Santos Silva and Tenreyro (2006) shows that the PPML is robust in the presence of heteroskedasticity. So I also try this alternative method.

Table 6 shows the results of using different methods to estimate the effects on language similarity on trade. Column 1 is the result of a log-linear regression which omits observations with zero trade flows, Column 2 shows the result of adding one (dollar) to all exports values before taking logs. Column 3 and 4 show the estimation results using Tobit and PPML estimator, respectively. From these results we can see that using OLS the effect of language similarity is positive and significant. And note that using either Tobit or PPML, the estimates of the effects of some key factors (such as the importer GDP and distance) become insignificant. Therefore, we can conclude that the Tobit and PPML estimators are not suitable in the dataset here, and in the following analysis I will use the OLS estimator to estimate the effect of language similarity.

3.2 Structural gravity model

Anderson and van Wincoop (2003) point out the importance of including the multilateral resistance terms in the gravity equation estimation, but their estimation strategy requires custom programming. Feenstra (2016) suggests that we can instead use fixed effects to account for the unobserved price indexes. So I try this fixed effects method, and the results are reported in Table 7. Column 1 to Column 3 are the same as their counterparts in the naïve gravity equations except including province-year and country-year fixed effects.

From the results, we can see that the effect of language similarity cannot be precisely estimated after controlling for province-quarter and country-quarter fixed effects. The coefficients all become negative, but note that when using OLS, either the effect of landlock or the effect of geographic distance cannot be precisely measured, and when using PPML, the effect of language similarity is not significant. Yotov et al (2016) also report that when using exporter and importer fixed effects, the estimates of the effects of common language and colonial ties become smaller, and when using PPML, the effect of colonial ties becomes negative and marginally significant, which is also not reasonable. As a result, I still use the naive gravity equation in the subsequent studies.

3.3 Robustness checks

3.3.1 Inter-province migration

When calculating the population share of people speaking a certain dialect, if in a certain province there are no any counties recorded as speaking this dialect according to the data of Liu et al. (2007), then the population share of people speaking this dialect in this province is assumed to be 0%. However, this value is underestimated due to the existence of inter-province migration. But because the 5 dialects of interest are mainly spoken in Southeast China, and as discussed in Section 2, most of the inter-province migration flows in China are from middle/western provinces to eastern provinces, the actual population shares which are assumed to be 0% are still very small. Thus, to examine the effect of this underestimation of population shares in some provinces, I reset those values from 0% to 1% and get another value of the language similarity index. Then I estimate the effect this second version of language similarity index, and the result is recorded in Table 8.

From Table 8, we can see that language similarity still has a positive and significant effect on trade. And the new coefficients are larger, which means that this effect is even slightly underestimated with the main version of the index.

3.3.2 Dialects and sub-dialects

In the 5 dialects of interest, Teochew and Hokien are in fact two subdialects of Minnan dialect (but they are used separately in classification of ethnic groups in Southeast Asian countries), while the other three dialects (Cantonese, Hakka and Hainanese) are different dialects themselves. To examine whether the similarity between Teochew and Hokien will affect the estimation result, I treat them as the same dialect group Minnan and calculate another version of language similarity index, following the same method described in Section 2 except that now there are only 4 dialects of interest (Cantonese, Hakka, Hainanese and Minnan). Then I estimate the effect this third version of language similarity index, and the result is recorded in Table 9.

From Table 9, we can see that when Teochew and Hokien are treated as the same dialect group Minnan, the overall language similarity still has a positive and significant effect on trade, though the effect becomes smaller.

3.3.3 Different ways of constructing indexes

I also try two other ways of constructing a language similarity index. One is to define

$$\text{languageSimilarity}_{\text{ij}}^{2} == \begin{cases} 1, if\ there\ are\ dialects\ commonly\ spoken\ in\ i\ and\ j \\ 0, otherwise \end{cases},$$

which is similar to the traditional way of measuring language similarity between two countries.

And another language similarity index is defined as the following:

$$\text{languageSimilarity}_{\text{ij}}^{3} = \sum_{d} \min \{popShare_{di}, popShare_{dj}\}$$

where $popShare_{dk}$ is the percentage of people speaking a certain dialect d in economy k, and $d \in \{\text{Cantonese}, \text{Teochew}, \text{Hakka}, \text{Hokkien}, \text{Hainanese}\}$.

Table 10 shows the results of estimating the effects of language similarity using the three indexes, and we can see that all three effects are positive, and using the first similarity index (which is the main version) the effect can be more precisely estimated.

4. Mechanisms

In this section, I will take a closer look at the possible channels through which language-based ethnic group identity influences trade.

4.1 Direct information cost

When trading with people speaking the same language, the language similarity can directly help reduce communication cost. Lohmann (2011) argued that when two countries do not speak the same language, speaking more similar languages should facilitate trade between them, since it reduces the cost of communication. Melitz (2008) also showed that a common language can facilitate trade through enabling people to communicate directly.

But this channel of directly reducing communication cost is not likely to play a big role in the region of Southeast Asia. There are two reasons. The first reason is that apart from speaking different dialects as their mother tongues, many people in China can speak Mandarin since it is the official language of China and is widely used in school education and media programs. According to a report from the Ministry of Education of PRC in 2020, more than 80% of Chinese can speak Mandarin, and this number would be higher for those businessmen/businesswomen involving in trade. And with the rise of China's economy in recent decades, there is an upsurge of learning Mandarin among Chinese in Southeast Asian countries. For example, in 1979 (the same year when China started to implement the "Reform and Opening Up" policy), the government of Singapore launched the Speak Mandarin Campaign to encourage the Singaporean Chinese to speak Mandarin, and according to Singapore Census of Population (2010), about 50% of Singaporean Chinese speak Mandarin frequently at home. So, for people who speak different Chinese dialects, many of them can still communicate directly via Mandarin (or even English). And the second reason is that although different Chinese dialects are mutually unintelligible, the written versions of them are with the same Chinese characters, and many contracts are written in Chinese or English where dialects do not play a role here. Therefore, the channel of directly reducing communication cost is not that important for trade among people who speak different Chinese dialects.

If the importance of this direct channel is greatly undermined because of the prevalence of another common language (Mandarin) or the same written version, then why the dialect similarity can still have a significantly positive effect on trade? The key is that many business negotiations occur in face-to-face communication, and that Mandarin is a second language to most Chinese people while those different dialects are their mother tongues. Nelson Mandela once said, "If you talk to a man in a language he understands, that goes to his head. If you talk to him in his own language, that goes to his heart." In fact, not all languages are equal; one's mother tongue is much important than other languages acquired later, because it plays an important role in shaping one's ethnic group identity. Hence, to see the whole picture, we need to look at this problem from the broader angle of ethnic group identity and examine the subsequent indirect effects it may bring about.

4.2 Indirect information cost

To enter export market, a firm needs to pay a fixed export cost, which includes the information cost, such as marketing research cost, search cost and verification cost.

4.2.1 Marketing research cost

People in the same ethnic group share similar cultural backgrounds and hence are more familiar with each other, which can reduce the fixed cost needed to learn about a foreign market before entering it and hence facilitate trade. Morales, Sheu and Zahler (2019) indicate that sharing similarities with a previous export destination has a large effect in reducing the cost of foreign market entry.

If with firm-level trade data, we can examine this channel through looking at how language similarity influences the extensive margins of trade.

4.2.2 Search cost

For people in the same ethnic group, they have more connections with each other and hence have a stronger network, which can help reduce the search cost of trade participants when searching for appropriate buyers or sellers. Rauch and Trindade (2002) showed that social networks have a considerable quantitative impact on international trade by helping to match buyers and sellers. Burchardi, Chaney and Hassan (2019) also showed that ethnic connections have a positive effect on FDI and trade via the channel of reducing information cost. And the increase of FDI will lead to more multinational firms, which will also indirectly facilitate trade.

4.2.3 Verification cost

People in the same ethnic group tend to trust more with each other, which can also facilitate trade by reducing verification costs. Guiso, Sapienza and Zingales (2009) found that the commonality between the two languages have a significant effect on bilateral trust and further showed that trust has a positive and economically and statistically significant effect on trade, FDI, and portfolio investment.

If with product-level trade data, we can test this verification cost reduction mechanism and the search cost reduction mechanism together, through looking at whether language similarity has a larger effect on trade of relationship-specific goods than that on trade of non-relationship-specific goods. But it is still a challenge to separate these two mechanisms.

4.3 Similar tastes

Apart from reducing information cost, people in the same ethnic group tend to have similar tastes, which also leads to more trade among them. If the formation of taste for an ethnic group is related to the abundant resources of their homeland (i.e., if their homeland has a comparative advantage of producing a certain type of goods, and hence people living there produce more and consume more and then gradually form a preference for that type of goods) , and if this taste persists after some people emigrates and the new country does not have a comparative advantage of producing that goods, then these people will tend to import this goods from their homeland. Atkin (2013) argues that adult taste favors the foods they consumed as a child and this taste persists over time. Zhang (2020) shows that immigrants' tastes drive them to import goods from those immigrants' countries of origin.

If with product-level trade data, we can examine this mechanism through looking at whether language similarity has a larger effect on trade of foods than that on trade of other goods such as manufacturing goods.

4.4 Social identity

There is another possible channel through which social group identity influences trade. Social identity is a person's sense of who they are based on their group membership. Tajfel and Turner (1979) proposed that the groups which people belonged to are an important source of pride and self-esteem, and there are also dissonance costs of group membership due to personal distance from average group member. And this psychological theory can be used in trade analysis. For example, Grossman and Helpman (2020) introduced social identity into a political-economics model of tariff formation.

In the setting with social identity, the utility function of a representative individual contains two components: material utility and psychological utility. When people trade with others in another economy but in the same social group, because of the dissonance cost in the psychological utility, exporting firms in the economy with higher living standards would choose to charge a lower markup to the importer so as to share some gains with people in the importing economy. This could increase the utility of people in the export economy by sacrificing some material utility but at the same time lowering the dissonance cost. This could also increase the utility of people in the import economy by increasing their material utility and lowering the dissonance cost. Similarly, customers in the economy with higher living standards would accept a higher markup charged by the exporting firms in the other economy. Hence, a common social identity can promote trade between people two economies (though the increase in trade is partly the diversion from the trade with other economies whose people do not share this social identity.) And we can have the following three predictions:

Prediction 1: If considering the fact that people in each economy pair share some common identities I, such as they are all Asians, then the difference in standard of living between the two economies will have a positive effect on trade since trade can bring gains to the two economies involved and help reduce the dissonance costs due to the difference in welfare of people in both sides.

I test prediction 1 using the following equation:

$$lnX_{ijt} = \beta_0 + \beta_1 lnGDP_{it} + \beta_2 lnGDP_{jt} + \beta_3 lnPop_{it} + \beta_4 lnPop_{jt} + \beta_5 lnDist_{ij} + \beta_6 commonBorder_{ij} + \beta_7 landlock_{ij} + \beta_8 PGDPGap * 1 + \beta_9 I_t + u_{ijt} \quad (2)$$

and expect that $\beta_8 > 0$.

Prediction 2: If apart from the basic common social identities I, people in two economies also share another social identity I', then the difference in standard of living between these two economies will have a larger positive effect on trade than that between another two economies that do not share this common social identity I'.

I test prediction 2 using the following equation:

$$lnX_{ijt} = \beta_0 + \beta_1 lnGDP_{it} + \beta_2 lnGDP_{jt} + \beta_3 lnPop_{it} + \beta_4 lnPop_{jt} + \beta_5 lnDist_{ij} + \beta_6 commonBorder_{ij} + \beta_7 landlock_{ij} + \beta_8 PGDPGap * 1 + \beta_9 PGDPGap * languageSimilarity_{ij} + \beta_{10} I_t + u_{ijt} \quad (3)$$

and expect that $\beta_9 > 0$

Prediction 3: If apart from the basic common social identities I, people in two economies also share another two social identities I' and I'', then I' and I'' will have some substitutional effects with each other in their joint effects with the difference in standard of living between these two economies.

To test this prediction, in addition to the language-based ethnic group identification, I also use the official classification of ethnic groups in China. In this classification system, China has 56 ethnic groups, such as Han, Zhuang, Hui, Manchu and Uyghur. In these 56 ethnic groups, Han is the majority while the other 55 are minorities. Because most of the Chinese immigrants in Southeast Asian countries are from Southeast China, where the majority is Han people, the majority of Chinese in those countries are also Han people. Hence, similar to the construction of language similarity index, I use the product of population shares of Han people to construct a Han similarity index to represent the ethnic group similarity within a province-country pair:

$$Han\text{Similarity}_{\text{ij}} = popShare_i^{Han} * popShare_j^{Han}$$

which represents the probability of if you randomly pick a person from a province i in China, and randomly pick another person from a certain Southeast Asian country j, and both of them are Han people. And I test prediction 3 using the following equation:

$$\begin{aligned} lnX_{ijt} = {} & \beta_0 + \beta_1 lnGDP_{it} + \beta_2 lnGDP_{jt} + \beta_3 lnPop_{it} + \beta_4 lnPop_{jt} + \beta_5 lnDist_{ij} \\ & + \beta_6 commonBorder_{ij} + \beta_7 landlock_{ij} + \beta_8 PGDPGap * 1 + \beta_9 PGDPGap \\ & * languageSimilarity_{ij} + \beta_{10} PGDPGap * HanSimilarity_{ij} + \beta_{11} PGDPGap \\ & * languageSimilarity_{ij} * HanSimilarity_{ij} + \beta_{12} I_t + u_{ijt} \qquad (4) \end{aligned}$$

and expect that $\beta_9 > 0$, $\beta_{10} > 0$ and $\beta_{11} < 0$.

From Table 11, column 1 and 2, we can see that the difference in GDP per capita does have a positive and significant effect on trade. And from column 3 and 4, we can see that language similarity significantly magnifies the positive effect of the difference in GDP per capita, and column 5 and 6 further indicate that two different ethnic group identities have a substitution effect with each other in their joint effects with the difference in standard of living between these two economies. These results all show some evidence that a common social identity does promote trade.

If with firm-level data, we can further examine this social identity mechanism by looking at how dialect group identity together with the difference in living standards affect firms' decisions on charging different markups.

5. Conclusion

Language similarity has a positive and significant effect on trade. But language itself is only part of the whole story. To see the whole picture, we need to observe from the broader angle of social group. When people trade with "their own people" (i.e., people in the same social group), it is easier for them to communicate with each other (if they speak the same language), they know more about each other, they trust more with each other, they have similar tastes, and they are more willing to share the gain from trade with each other, all of which will promote trade between them.

There are several future directions for this work. First, I would like to develop a model based on the theory of social identity to better guide the empirical study. Second, I want to use firm-level and product-level data to further investigate the mechanisms through which social group identity influences trade. Finally, I hope to find out a better way to carry out the structural estimation.

Appendix A

I describe in this appendix data sources:

1. Trade data

The province-country pair trade data between 31 provinces of Mainland China and 8 Southeast Asian countries is from China Custom. The data are at a monthly frequency, and I aggregate them to quarterly data. The data are recorded in the unit of dollars.

2. Dialects data

The dialects of interest are Cantonese, Teochew, Hakka, Hokkien and Hainanese, as described in Section 2. There are two parts of dialects data needed: dialects spoken in Southeast Asian countries and dialects spoken in China provinces.

The population data of each dialect group of Singapore is from Singapore Census of Population (2010), and the dialect data for the other 7 Southeast Asian countries are from Li (2012).

The data sources of dialect data for each China province are a bit more various: the data for Guangdong are from Guangdong Local Gazetteers (2004), the data for Hainan are from Hainan Local Gazetteers (1994), the data about Hakka dialect spoken in Jiangxi, Fujian and Hunan is from Xie and Huang (2007). For the rest of the dialects spoken in other provinces, I get the data about the county-level distribution of dialects spoken in China from Liu et al. (2007), and then map this distribution with the China Census of Population (2010) data to get the population speaking each dialect in each province.

Except for the dialects data spoken in Hainan province, which is already recorded in percentage in the data source, I calculate the population share of dialects spoken in each economy (a province or a Southeast Asian country) through dividing the population of each dialect spoken in each economy by the total population of that economy in the corresponding year. And then I use this population share data in the period of interest (2017Q1-2019Q4), as described in Section 2.

3. GDP data

The quarterly nominal GDP data of the 8 Southeast Asian countries are from IMF's International Financial Statistics, except that the GDP data of Malaysia from 2019Q1 to 2019Q4 are from the Treasury of Malaysia.

The quarterly nominal GDP data of the 31 provinces of Mainland China are from the National Bureau of Statistics of China.

The GDP data mentioned above are all recorded in domestic currency, so I also use the exchange rate data to convert them to US dollars.

4. Exchange rate data

The quarterly exchange rate data are from IMF's International Financial Statistics.

5. Population data

The population data of the 8 Southeast Asian countries are from the World Bank's World Development Indicators. And the population data of the 31 provinces are from the National Bureau of Statistics of China. These data are at a yearly frequency, and I apply the linear interpolation method to estimate the quarterly population for each economy.

To calculate the population of each dialect spoken in a certain province of China, I also need city-level or even county-level population data. These more fine-grained population data are from the China Census of Population (2010).

6. Ethnic groups data

The data on Han ethnic group in China are from the China Census of Population (2010). And the specific corresponding data among the Chinese in Southeast Asian countries are not available. But because most of the Chinese immigrants in Southeast

Asian countries are from Southeast China, where the majority is Han people, I assume in this paper that the Chinese in those countries are all Han people.

7. Distance data

The distance data between the capital cities of each province-country pair are from distance-cities.com.

8. Other geographic data

The data about whether an economy is landlocked and whether a province-country pair has a common border are from Google Map.

References

Anderson, James E. and Eric Van Wincoop (2003): "Gravity with gravitas: A solution to the border puzzle," American economic review, 93, 170–192.

Atkin, David (2013): "Trade, Tastes and Nutrition in India." American Economic Review, 103, 1629–1663

Burchardi, Konrad B., Thomas Chaney, and Tareka A. Hassan (2019): "Migrants, Ancestors, and Foreign Investments," Review of Economic Studies, 86, 1448–1486

Chaney, Thomas (2008): "Distorted gravity: the intensive and extensive margins of international trade," American Economic Review 98 (4), 1707–21.

China Census of Population (2010), National Bureau of Statistics of China

Eaton, Jonathan, and Samuel S. Kortum (2002): "Technology, geography, and trade," Econometrica, 70, 1741–1779.

Fearon, James D. (2003): "Ethnic and Cultural Diversity by Country," Journal of Economic Growth, 8, 195–222.

Feenstra, Robert C. (2016): "Advanced International Trade: Theory and Evidence," Princeton, NJ: Princeton University Press.

Fernandes, Ana M., Peter J. Klenow, Sergii Meleshchuk, Martha Denisse Pierola, and Andres Rodríguez-Clare (2018): "The Intensive Margin in Trade," (Working Paper No. 25195). National Bureau of Economic Research.

Grossman, Gene Michael, and Elhanan Helpman, (2020): "Identity Politics and Trade Policy," The Review of Economic Studies.

Guangdong Local Gazetteers (2004), (广东省志), The Office of Guangdong Chronicles, ISBN 7218044077. (in Chinese)

Guiso, Luigi, Paola Sapienza and Luigi Zingales (2009): "Cultural biases in economic exchange," Quarterly Journal of Economics, 124, 1095–1131.

Hainan Local Gazetteers (1994), (海南省志), The Office of Hainan Chronicles, ISBN 7544200787. (in Chinese)

Helpman, Elhanan, Marc Melitz, and Yona Rubinstein (2008): "Estimating Trade Flows: Trading Partners and Trading Volumes," Quarterly Journal of Economics 123 (2), 441–487.

Li, Yong (2012): "Language, History, Boundary: The Transition of Chinese Ethnic Relationship in Southeast Asian,"(语言、历史、边界：东南亚华人族群关系的变迁), Editor-in-Chief, Qiu Jin, "Blue Book of Overseas Chinese: Annual Report on Overseas Chinese Study,"(华侨华人蓝皮书：华侨华人研究报告), China: Social Sciences Academic Press, 122-155. (in Chinese)

Li, Yang, Hui Liu, and Qing Tang (2015): "Spatial-temporal patterns of China's interprovincial migration during 1985-2010," (1985-2010 年中国省际人口迁移时

空格局特征), Geographical Research, 2015, 34(6): 1135-1148. (in Chinese)

Liu, Yuyun, Xianxiang Xu, and Zekai Xiao (2015): "The Pattern of Labor Cross-dialects Migration," (劳动力跨方言流动的倒 U 型模式), Economic Research Journal, Vol 50, No. 10, Pages 134-146+162. (in Chinese)

Lohmann, Johannes (2011): "Do language barriers affect trade?", Economics Letters, 110: 2, February, pp. 159-162.

Mai, Liu Fang (1985): "Dialect group identity: a study of Chinese subethnic group in early Malaya," (方言群認同：早期星馬華人的分類法則), Institute of Ethnology, Academia Sinica, Monograph series B no.14. (in Chinese)

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

Morales, Eduardo, Gloria Sheu and Andres Zahler (2019): "Extended Gravity," Review of Economic Studies, 86(6): 2668–2712.

Rauch, James E. and Vitor Trindade (2002): "Ethnic Chinese Networks In International Trade," The Review of Economics and Statistics, MIT Press, vol. 84(1), pages 116-130, February.

Santos Silva, J. M.C., and Silvana Tenreyro (2006): "The log of gravity," The Review of Economics and Statistics 88(4), 641–658.

Singapore Census of Population (2010), Singapore Department of Statistics, 44-45

Tinbergen, Jan (1962): "Shaping theWorld Economy: Suggestions for an International Economic Policy," Twentieth Century Fund, New-York.

Tajfel, Henri and John Turner (1979: "An Integrative Theory of Intergroup Conflict", in Austin, W. G. and Worchel, S. (eds) The Social Psychology of Intergroup Relations (Monterey: Brooks/Cole Pub. Co.).

Xie, Liuwen, and Xuezhen Huang (2007): "Classification/Distribution of Hakka Group," (客家方言的分区), Dialect, 2007(03). (in Chinese)

Xu, Bao Hua and Ichiro Miyata (1999): "The Great Dictionary of Chinese Dialects," (汉语方言大词典), Beijing: Zhonghua Book Company. (in Chinese)

Yotov, Yoto V., Roberta Piermartini, José-Antonio Monteiro and Mario Larch (2016): "An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model," WTO.

Zhang, Penglong (2020): "Home-biased gravity: The role of migrant tastes in international trade," World Development, 129.

Zhuang, Guotu (2008): "On the Four waves : history of Chinese migration into Southeast Asia," (论中国人移民东南亚的四次大潮), Southeast Asian Affairs, No. 1 2008, General Serial No. 133. (in Chinese)

Tables and Figures

Table A1
List of Provinces of Mainland China

Anhui	Hainan	Jiangxi	Shanxi
Beijing	Hebei	Jilin	Sichuan
Chongqing	Heilongjiang	Liaoning	Tianjin
Fujian	Henan	Ningxia	Tibet
Gansu	Hubei	Qinghai	Xinjiang
Guangdong	Hunan	Shaanxi	Yunnan
Guangxi	Inner Mongoria	Shandong	Zhejiang
Guizhou	Jiangsu	Shanghai	

Notes: This list does not include Hongkong, Macao, and Taiwan because the political systems and tariff policies of these three economies are different than that of those other 31 provinces of Mainland China and the quarterly trade data between these three economies and other countries are not available.

Table A2
List of Southeast Asian Countries

Brunei	Indonesia	Philippines	Thailand
Burma	Malaysia	Singapore	Vietnam

Notes: In fact, there are 2 more Southeast Asian countries Cambodia and Laos, but the quarterly GDP data of these two countries are not available, so I do not take them into account.

Table A3
Main Variables

Variable	Description
export_value	bilateral export value, in dollars
province_GDP	nominal GDP of a province, in dollars
country_GDP	nominal GDP of a country, in dollars
province_pop	population of a province
country_pop	population of a country
distance	distance between the capital cities of the two economy, in kilometers
PGDP_gap	the gap of the GDP per capita within a province-country pair
landlock	a binary dummy indicating whether one economy of the province-country pair is landlocked
common_border	a binary dummy indicating whether a province and a country share a common border

Table A4

Summary Statistics

	mean	sd	min	max
export_value	**3.350e+08**	**7.244e+08**	**1.000**	**1.014e+10**
exporter_GDP	**1.078e+11**	**8.213e+10**	**2.895e+09**	**4.291e+11**
importer_GDP	**1.053e+11**	**8.417e+10**	**2.895e+09**	**4.291e+11**
exporter_population	**64906415.966**	**61754879.837**	**420975.000**	**2.706e+08**
importer_population	**64675118.368**	**63383568.833**	**420975.000**	**2.706e+08**
distance_km	**3041.662**	**1159.297**	**326.410**	**6101.260**
PGDP_gap	**3372.807**	**4312.479**	**0.831**	**15733.837**
ln_export_value	**17.341**	**3.071**	**0.000**	**23.040**
ln_exporter_GDP	**25.062**	**0.954**	**21.786**	**26.785**
ln_importer_GDP	**24.947**	**1.134**	**21.786**	**26.785**
ln_exporter_pop	**17.517**	**1.154**	**12.950**	**19.416**
ln_importer_pop	**17.365**	**1.469**	**12.950**	**19.416**
ln_distance	**7.931**	**0.460**	**5.788**	**8.716**
ln_PGDP_gap	**7.397**	**1.286**	**0.605**	**9.664**
landlock	**0.653**	**0.476**	**0.000**	**1.000**
common_border	**0.013**	**0.115**	**0.000**	**1.000**
Language_similarity	**0.003**	**0.013**	**0.000**	**0.147**

Table A5

Correlation Table

	ln_distance	common_border	landlock	languageSimilarity
ln_distance	**1**			
common_border	**-0.388*** **	**1**		
landlock	**0.067*** **	**0.076*** **	**1**	
languageSimilarity	**-0.045*** **	**-0.013**	**-0.191*** **	**1**

Notes: * p<0.05 ** p<0.01 *** p<0.001

Table 1
Basic OLS estimation results

	(1)	(2)	(3)	(3a)
				Betas
VARIABLES	ln(X)	ln(X)	ln(X)	ln(X)
ln_exporter_GDP	1.184***	1.053***	1.024***	0.390***
	(0.107)	(0.107)	(0.108)	(0.032)
ln_importer_GDP	0.937***	0.679***	0.652***	0.248***
	(0.073)	(0.088)	(0.088)	(0.032)
ln_distance	-0.934***	-0.647***	-0.596***	-0.087***
	(0.195)	(0.201)	(0.203)	(0.031)
common_border	3.644***	3.699***	3.733***	0.133***
	(0.597)	(0.554)	(0.547)	(0.029)
landlock	-1.952***	-2.067***	-1.994***	-0.304***
	(0.196)	(0.191)	(0.195)	(0.028)
ln_exporter_pop		0.219*	0.254**	0.123***
		(0.124)	(0.127)	(0.040)
ln_importer_pop		0.367***	0.396***	0.192***
		(0.075)	(0.076)	(0.037)
languageSimilarity			17.975***	0.073***
			(5.876)	(0.028)
Observations	5,402	5,402	5,402	5,402
economy pairs	479	479	479	479
quarter FE	YES	YES	YES	YES

Notes: Column 3a reports standardized coefficients from Column 3. Robust standard errors are shown in parenthesis (***p<0.01, ** p<0.05, * p<0.1).

Table 2
Probit estimation results

VARIABLES	(1) Marginal effects T=1	X>0	(1a) Standardized marginal effects T=1	X>0
ln_exporter_GDP	1.200***	1.403***
	(0.184)	(0.215)
ln_importer_GDP	0.991***	1.159***
	(0.205)	(0.239)
ln_exporter_pop	0.453***	0.673***
	(0.153)	(0.227)
ln_importer_pop	-0.018	-0.027
	(0.155)	(0.231)
ln_distance	-0.525	-0.236
	(0.528)	(0.237)
landlock	-2.268***	-1.060***
	(0.581)	(0.272)
dialect_pct_product_sum	84.933	1.062
	(65.747)	(0.822)
Observations	5,880	5,880
economy pairs	490	490
quarter FE	YES	YES

Notes: The common_border variable is dropped in the estimation because it predicts success perfectly. Column 1 shows the result of Probit estimation. Column 1a standardizes Column 1 to report the probability change from one standard deviation increase in the independent variables. Robust standard errors are shown in parenthesis (***p<0.01, ** p<0.05, * p<0.1).

Table 3
Zero export observations grouped by province-country pair

province	GDP_rank_R	country	GDP_rank_R	zero count
Tibet	**1**	Brunei	**1**	11
Tibet	**1**	Burma	**2**	8
Tibet	**1**	Viet Nam	3	8
Tibet	**1**	Malaysia	4	2
Tibet	**1**	Singapore	5	2
Tibet	**1**	Philippines	6	4
Tibet	**1**	Thailand	7	6
Tibet	**1**	Indonesia	8	6
Qinghai	**2**	Brunei	**1**	6
Qinghai	**2**	Burma	2	5
Qinghai	**2**	Philippines	6	1
Ningxia	3	Brunei	**1**	3
Hainan	4	Brunei	**1**	3
Gansu	5	Brunei	**1**	6
Jilin	6	Brunei	**1**	3
Xinjiang	7	Brunei	**1**	3
Guizhou	10	Brunei	**1**	2
--	--	--	**--**	Total: 79

Notes: This table only contains information about the zero observations in the export flows. The "GDP_rank_R" column is the ranking of GDP in REVERSE order ranking for each economy (in 31 provinces or 8 countries). And the "zero count" column is the number of zeros that occur for each province-country pair. The maximum count is 12 since there are only 12 quarters of observations for each province-country pair.

Table 4

Small trade (X< $100) observations

province	GDP_rank_R	country	GDP_rank_R	exp/imp	quarter	trade value ($)
Zhejiang	28	Brunei	**1**	import	2019q2	1
Zhejiang	28	Brunei	**1**	import	2017q4	3
Hubei	25	Brunei	**1**	import	2017q2	3
Guizhou	10	Brunei	**1**	export	2019q4	4
Zhejiang	28	Brunei	**1**	import	2017q2	5
Hubei	25	Brunei	**1**	import	2019q3	6
Jilin	6	Brunei	**1**	export	2019q3	10
Jilin	6	Brunei	**1**	export	2019q4	16
Jilin	6	Brunei	**1**	export	2019q1	21
Zhejiang	28	Brunei	**1**	import	2019q3	23
Guangxi	13	Brunei	**1**	import	2017q4	30
Shaanxi	18	Burma	**2**	import	2019q3	43
Heilongjiang	8	Burma	**2**	import	2018q4	69
Qinghai	**2**	Burma	**2**	export	2019q3	76
Tianjin	9	Brunei	**1**	import	2019q3	82
Hubei	25	Burma	**2**	import	2019q1	87
Shandong	29	Brunei	**1**	import	2018q3	90

Notes: The "GDP_rank_R" column is the ranking of GDP in REVERSE order ranking for each economy (in 31 provinces or 8 countries). And the trade values are in the unit of dollars.

Table 5
Exports with different measurement units

	(2)	(3)	(4)
	OLS	OLS	OLS
VARIABLES	ln(X+1)	ln(X/1000+1)	ln(X/1000000+1)
ln_exporter_GDP	1.459***	1.013***	0.482***
	(0.143)	(0.093)	(0.048)
ln_importer_GDP	0.961***	0.679***	0.315***
	(0.147)	(0.094)	(0.044)
ln_exporter_pop	1.001***	0.622***	0.271***
	(0.173)	(0.105)	(0.051)
ln_importer_pop	0.522***	0.485***	0.341***
	(0.121)	(0.085)	(0.047)
ln_distance	-0.581*	-0.446*	-0.304*
	(0.315)	(0.230)	(0.159)
common_border	4.619***	4.182***	3.405***
	(0.774)	(0.582)	(0.412)
landlock	-2.670***	-2.357***	-1.865***
	(0.275)	(0.208)	(0.148)
languageSimilarity	28.591***	23.372***	16.031***
	(8.186)	(6.450)	(5.037)
Observations	5,952	5,952	5,952
province-country pairs	496	496	496
quarter FE	YES	YES	YES

Notes: Column 1, 2 and 3 show the results of the method of adding one to observed exports before taking logs, but with different measurement units (the units change from dollar to thousands to millions respectively). Robust standard errors are shown in parenthesis (***p<0.01, ** p<0.05, * p<0.1).

Table 6

The effect of language similarity on trade

	(1)	(2)	(3)	(4)
	OLS	OLS	Tobit	PPML
VARIABLES	ln(X)	ln(X+1)	ln(X+1)	X/1000000
ln_exporter_GDP	1.024***	1.459***	1.285***	0.323***
	(0.108)	(0.143)	(0.179)	(0.121)
ln_importer_GDP	0.652***	0.961***	0.873***	0.137
	(0.088)	(0.147)	(0.169)	(0.085)
ln_exporter_pop	0.254**	1.001***	1.425***	0.938
	(0.127)	(0.173)	(0.167)	(0.596)
ln_importer_pop	0.396***	0.522***	0.632***	0.842
	(0.076)	(0.121)	(0.160)	(0.590)
ln_distance	-0.596***	-0.581*	-0.447	1.360
	(0.203)	(0.315)	(0.449)	(1.466)
common_border	3.733***	4.619***	4.757***	4.196***
	(0.547)	(0.774)	(1.756)	(1.486)
landlock	-1.994***	-2.670***	-2.957***	-1.507***
	(0.195)	(0.275)	(0.396)	(0.271)
languageSimilarity	17.975***	28.591***	35.110**	17.325**
	(5.876)	(8.186)	(14.625)	(7.704)
Observations	5,402	5,952	5,952	5,952
economy pairs	479	496	496	496
quarter FE	YES	YES	YES	YES

Notes: Column 1 is the result of a log-linear regression which omits observations with zero trade flows, Column 2 shows the result of adding one (dollar) to all exports values before taking logs. Column 3 shows the result of Tobit estimation. Column 4 show the estimation result using PPML. Robust standard errors are shown in parenthesis (***p<0.01, ** p<0.05, * p<0.1)

Table 7
Structural gravity models

VARIABLES	(1) OLS ln(X)	(2) OLS ln(X+1)	(3) PPML X/1000000
ln_distance	-1.039***	-0.562	-0.919***
	(0.383)	(0.578)	(0.308)
common_border	3.396***	4.346***	4.258***
	(1.095)	(1.473)	(0.882)
landlock	-1.252	-3.122*	-1.737**
	(0.920)	(1.651)	(0.838)
languageSimilarity	-2.811	-8.772*	-1.754
	(3.098)	(4.544)	(2.427)
Observations	5,402	5,952	5,952
province-country pairs	479	496	496
province-quarter FE	YES	YES	YES
country-quarter FE	YES	YES	YES

Notes: Column 1 is the result of a log-linear regression which omits observations with zero trade flows, Column 2 shows the result of adding one (dollar) to all exports values before taking logs, and Column 3 shows the estimation result using PPML. Robust standard errors are shown in parenthesis (***$p<0.01$, ** $p<0.05$, * $p<0.1$).

Table 8
Robustness checks for inter-province migration

VARIABLES	(1) OLS ln(X)	(2) OLS ln(X+1)	(3) OLS ln(X)	(4) OLS ln(X+1)
ln_exporter_GDP	1.024***	1.459***	1.000***	1.414***
	(0.108)	(0.143)	(0.108)	(0.143)
ln_importer_GDP	0.652***	0.961***	0.627***	0.916***
	(0.088)	(0.147)	(0.088)	(0.147)
ln_exporter_pop	0.254**	1.001***	0.288**	1.053***
	(0.127)	(0.173)	(0.129)	(0.174)
ln_importer_pop	0.396***	0.522***	0.425***	0.574***
	(0.076)	(0.121)	(0.077)	(0.123)
ln_distance	-0.596***	-0.581*	-0.597***	-0.575*
	(0.203)	(0.315)	(0.203)	(0.314)
common_border	3.733***	4.619***	3.720***	4.603***
	(0.547)	(0.774)	(0.545)	(0.770)
landlock	-1.994***	-2.670***	-1.978***	-2.630***
	(0.195)	(0.275)	(0.195)	(0.273)
languageSimilarity	17.975***	28.591***		
	(5.876)	(8.186)		
languageSimilarity_v2			24.661***	41.775***
			(7.829)	(12.250)
Observations	5,402	5,952	5,402	5,952
province-country pairs	479	496	479	496
quarter FE	YES	YES	YES	YES

Notes: Column 3 and 4 are the results for the second version of language similarity index while Column 1 and 2 are for the original one. Robust standard errors are shown in parenthesis (***p<0.01, ** p<0.05, * p<0.1).

Table 9
Robustness checks for dialects and sub-dialects

VARIABLES	(1) OLS ln(X)	(2) OLS ln(X+1)	(3) OLS ln(X)	(4) OLS ln(X+1)
ln_exporter_GDP	1.024***	1.459***	1.043***	1.486***
	(0.108)	(0.143)	(0.108)	(0.142)
ln_importer_GDP	0.652***	0.961***	0.670***	0.988***
	(0.088)	(0.147)	(0.088)	(0.147)
ln_exporter_pop	0.254**	1.001***	0.232*	0.972***
	(0.127)	(0.173)	(0.125)	(0.171)
ln_importer_pop	0.396***	0.522***	0.377***	0.494***
	(0.076)	(0.121)	(0.075)	(0.120)
ln_distance	-0.596***	-0.581*	-0.639***	-0.649**
	(0.203)	(0.315)	(0.201)	(0.313)
common_border	3.733***	4.619***	3.695***	4.557***
	(0.547)	(0.774)	(0.553)	(0.782)
landlock	-1.994***	-2.670***	-2.034***	-2.726***
	(0.195)	(0.275)	(0.195)	(0.275)
languageSimilarity	17.975***	28.591***		
	(5.876)	(8.186)		
languageSimilarity_v3			3.003**	5.513***
			(1.297)	(1.854)
Observations	5,402	5,952	5,402	5,952
province-country pairs	479	496	479	496
quarter FE	YES	YES	YES	YES

Notes: Column 3 and 4 are the results for the third version of language similarity index while Column 1 and 2 are for the original one. Robust standard errors are shown in parenthesis (***p<0.01, ** p<0.05, * p<0.1).

Table 10
Robustness checks for different language similarity indexes

VARIABLES	(1) OLS ln(X)	(2) OLS ln(X+1)	(3) OLS ln(X)	(4) OLS ln(X+1)	(5) OLS ln(X)	(6) OLS ln(X+1)
ln_exporter_GDP	1.024***	1.459***	1.017***	1.461***	1.017***	1.455***
	(0.108)	(0.143)	(0.109)	(0.142)	(0.108)	(0.143)
ln_importer_GDP	0.652***	0.961***	0.645***	0.963***	0.645***	0.957***
	(0.088)	(0.147)	(0.088)	(0.146)	(0.089)	(0.147)
ln_exporter_pop	0.254**	1.001***	0.244*	0.987***	0.256**	1.001***
	(0.127)	(0.173)	(0.125)	(0.171)	(0.127)	(0.173)
ln_importer_pop	0.396***	0.522***	0.396***	0.509***	0.401***	0.522***
	(0.076)	(0.121)	(0.075)	(0.120)	(0.076)	(0.121)
ln_distance	-0.596***	-0.581*	-0.362	-0.330	-0.593***	-0.586*
	(0.203)	(0.315)	(0.227)	(0.351)	(0.203)	(0.314)
common_border	3.733***	4.619***	3.931***	4.836***	3.744***	4.630***
	(0.547)	(0.774)	(0.538)	(0.789)	(0.546)	(0.774)
landlock	-1.994***	-2.670***	-1.985***	-2.695***	-1.989***	-2.680***
	(0.195)	(0.275)	(0.192)	(0.273)	(0.195)	(0.276)
languageSimilarity	17.975***	28.591***				
	(5.876)	(8.186)				
languageSimilarity_2			0.858***	0.994***		
			(0.206)	(0.339)		
languageSimilarity_3					7.384***	10.229***
					(1.925)	(2.661)
Observations	5,402	5,952	5,402	5,952	5,402	5,952
province-country pairs	479	496	479	496	479	496
quarter FE	YES	YES	YES	YES	YES	YES

Notes: This table shows the results of using different language similarity indexes. Robust standard errors are shown in parenthesis (***p<0.01, ** p<0.05, * p<0.1).

Table 11
Social identity

VARIABLES	(1) OLS ln(X)	(2) OLS ln(X+1)	(3) OLS ln(X)	(4) OLS ln(X+1)	(5) OLS ln(X)	(6) OLS ln(X+1)
ln_exporter_GDP	1.039***	1.480***	1.018***	1.445***	0.723***	0.922***
	(0.109)	(0.140)	(0.109)	(0.142)	(0.110)	(0.142)
ln_importer_GDP	0.665***	0.980***	0.645***	0.946***	0.353***	0.418***
	(0.086)	(0.142)	(0.086)	(0.143)	(0.085)	(0.138)
ln_exporter_pop	0.252**	1.004***	0.277**	1.039***	0.663***	1.592***
	(0.125)	(0.170)	(0.127)	(0.172)	(0.138)	(0.171)
ln_importer_pop	0.401***	0.526***	0.422***	0.561***	0.739***	1.117***
	(0.075)	(0.118)	(0.076)	(0.120)	(0.086)	(0.139)
ln_distance	-0.658***	-0.677**	-0.620***	-0.612*	-0.961***	-1.188***
	(0.199)	(0.312)	(0.201)	(0.313)	(0.204)	(0.309)
common_border	3.667***	4.517***	3.693***	4.564***	3.183***	3.696***
	(0.549)	(0.779)	(0.544)	(0.770)	(0.558)	(0.787)
landlock	-2.026***	-2.733***	-1.978***	-2.649***	-2.107***	-2.813***
	(0.191)	(0.273)	(0.194)	(0.275)	(0.187)	(0.265)
ln_PGDP_gap	0.083**	0.118**	0.077**	0.108**	-0.034	-0.090*
	(0.033)	(0.050)	(0.033)	(0.050)	(0.034)	(0.053)
ln_PGDP_gap * languageSimilarity			1.545***	2.587***	4.432**	7.868**
			(0.549)	(0.788)	(1.781)	(3.117)
ln_PGDP_gap * HanSimilarity					0.622***	1.103***
					(0.065)	(0.089)
ln_PGDP_gap * languageSimilarity * HanSimilarity					-8.062***	-14.834***
					(2.812)	(4.970)
Observations	5,402	5,952	5,402	5,952	5,402	5,952
province-country pairs	479	496	479	496	479	496
quarter FE	YES	YES	YES	YES	YES	YES

Notes: Column 1, 3 and 5 are the results of log-linear regressions which omit observations with zero trade flows, Column 2, 4 and 6 show the result of adding one (dollar) to all exports values before taking logs. Robust standard errors are shown in parenthesis (***p<0.01, ** p<0.05, * p<0.1).