For office use only	Team Control Number	For office use only
T1 ________________	**91566**	F1 ________________
T2 ________________		F2 ________________
T3 ________________	Problem Chosen	F3 ________________
T4 ________________	**B**	F4 ________________

2018
MCM/ICM
Summary Sheet

Forecasting the Language Distribution

Summary

Language shift is a common and complicated phenomenon since ancient times, which has far-reaching influence on the development of the human civilization. The language shift is influenced by many factors. We study the geographically distributed language shift model and provide solutions based on language shift and economic benefits to the locations of new international offices for a service company.

To get the geographical language transformation model, we start off by creating a simple model, and then make it more sophisticated.

- **The Worldwide Language Shift Model** reflects the situation when two or more languages communicate. The model takes into account factors such as government support, tourism, international business relations and technological progress.
- **The Domestic Language Shift Model** is the application of the worldwide language shift model in country. We adjust some parameters to make this model fit the domestic situation better.
- **The Geographical Language Shift Model** is based on the domestic language shift model. This model contains the impact of migration on language shift.

We collect rich and effective data, fit the unknown parameters and get the geographical distribution of various languages in the future. Through the sensitivity analysis, we prove the stability and error tolerance of the model. By the model implementation, we found that the geographical language distribution remains basically stable in the next 50 years.

To address the locations of international offices, we take into account the predicted geographical language distribution and the economic need. In our opinion, choosing the location of an international office should make the company more profitable. To this end, we establish a model of the impact of geographical language distribution on the company's effectiveness. We identify the six cities most suitable for hosting international offices in both short and long terms are: Paris, Tokyo, Milan, Toronto, Bombay, Brussels. We also put forward our own ideas about the company's long-term plan.

In a word, we predict the language development for the next 50 years according to the geographical distribution language shift model, and provide effective reference for the locations of international offices.

Contents

To: Chief Operating Officer
From: Investigation Team
Date: 12 February, 2018
Subject: Suggestions for New Office Locations

We have completed an analysis of the locations of your company's international offices as well as the trends of global languages.

First, we set up a model of global language competition and summed up the rules of global language shift by collecting considerable and reliable data. We analyzed and predicted the geographical distribution of global languages in the next 50 years. Second, based on the changes of languages, the recruitment cost and the development level of the city where the office is located in, we provided 6 suitable office locations. Finally, we combined scientific and technological development factors, and gave reasonable suggestions to the long-term plan of setting up new offices in the future.

Based on our analysis, here are some suggestions:

- It is advisable to set up offices in the following six cities. The languages correspond to each office are as follows:

City	**Country**	**Corresponding language**
Paris	France	France, Spanish, English
Tokyo	Japan	Japanese, English
Milan	Italy	Italian, France, English
Toronto	Canada	English, French
Bombay	India	English, Hindustani, Bengali
Brussels	Belgium	France, Dutch,English

- We do not recommend frequent replacements of office sites, because there is a certain stability of the distribution of the world's languages, no major changes would occur within a short time. At the same time, replacing an office is costly, and adversely affecting the company's profits.
- Profits brought by language advantages will be reduced due to the rapid expansion of global communications. While the cost of employing may increase, the amount of existing offices may no longer be optimal. If you can provide more specific data, we will be able to give more reasonable advice on the long-term plans for the new offices.

Our modeling analysis report contains more detailed theoretical information. Please contact us promptly if there are any questions.

1 Introduction

1.1 Background

It is generally taken for granted that language, as a concomitant of culture, can spread. With the trend of globalization and the world's cultural exchange, language transfer and integration are also more common. Nowadays more and more people can speak two or even more languages.

The shift and spread of language can be seen through the amount of speakers, including native speakers plus second or third, etc. language speakers. However, the total number of speakers of a language fluctuates under the influence of various complicated factors. These factors involve political, economic, diplomatic, social relations and other aspects, such as government-mandated official languages, tourism among nations, migration and population movements, the promotion of new social media (facebook, Twitter, etc.) and so on.

1.2 Restatement of the problem

We are required to predict the spread and development of languages all over the world under the influence of several factors and help a large multinational service company to determine the locations of new offices.

The problem can be analyzed into three parts:

- Develop a model of the distribution of various language speakers over time based on impact factors and predict what will happen to the number of speakers of each language in the next 50 years.
- Use the model to predict the geographic distributions of languages in the next 50 years.
- Determine the locations of new international offices and the languages used in the new offices based on the modeling results.

1.3 Related Work

Mcmahon(1994) and Mufwene(2001) proposed that the languages change of a region is caused by a mechanism named language shift. Language shift is a process that takes place in the region where obsess more than one language. The members of the region abandon their initial language in favor of another. So, despite the migration. According to Abrams and Strogatz(2003), Anna and Roman(2010), Katharina and Gero(2017), language shift is modeled as a competition between two communities who use different languages, and the motivation of the language shift is to chase to better opportunity provided by another language. Abrams-Strogatz (A-S) Model is the most widely mathematic model of describing the changes in the patterns of language shift between two language communities. The A-S Model shows the temporal population shift of both languages, which results in

$$\frac{dn_A}{dt} = n_B P_{B\to A}(n_A, s_A) - n_A P_{A\to B}(n_B, s_B) \quad (1)$$

$$\frac{dn_B}{dt} = -\frac{dn_A}{dt} = n_A P_{A\to B}(n_B, s_B) - n_B P_{B\to A}(n_A, s_A) \quad (2)$$

where n_A and n_B represent the proportion of each language (language A and language B) of the total population. $P_{B\to A}(n_A, s_A)$ is the possibility that an individual from language B's community shifts to language B's community shifts, in other words, $P_{B\to A}(n_A, s_A)$ is the shift rate from B to A. The calculation method of $P_{B\to A}(n_A, s_A)$ is defined as

$$P_{B\to A}(n_A, s_A) = c n_A^a s_A \quad (3)$$

where c is the maximum shift rate, a quantifies the resistance level of language B speakers to change their language to A. So, n_A^a is a factor that measures the consistency of the attractiveness and language community size of A. s_A is the language status, which represents the social and economic opportunities afforded to the speakers of language A relative to language B. The higher the proportion of a language, the higher shift rate and its status and the lower the resistance, the higher its attractiveness and therefore speakers of other languages are more likely to use this language in the future.
In the same way, $P_{B\to A}(n_A, s_A)$ is defined as

$$P_{A\to B}(n_B, s_B) = c n_B^a s_B \quad (4)$$

As s_A and s_B are two opposite factors, we have $s_A + s_B = 1$.
Pinasco and Pomanelli(2006) developed an expanded model of A-S Model which takes the natural reproduction of language communities into account, and the natural reproduction of language A is

$$NR_A = r_A n_A (1 - n_A/K_A) \quad (5)$$

where r_A is the maximum natural growth rate and K_A is carrying capacity of language A in this region.

2 Assumptions and Notations

2.1 Assumptions

To simplify our problems, we make the following basic assumptions, each of which is properly justified.

- **Use 26 languages on behalf of all languages.** Because this 26 languages have 50 million or more total speakers and have a large impact all over the world. The languages are listed in the appendix.
- **Use 36 countries on behalf of all countries.** The update time of language data from *Ethnologue* is usually slow. We can not collect data about all countries all languages by time. We have access to data series over time in major languages of the world (26) and cross-section data updated in 2016 for countries in various countries in the world. Considering the company as a service company, in the selection of the

international office location, the population and economic level of a country also have some requirements. Therefore, we mainly study the top 30 countries in the world in terms of population ranking or the top 30 countries in the world in terms of GDP. We think these countries are the major influencers of linguistic changes and linguistic exchanges and major players in global economy. In addition, we remove countries lacking of relevant data (for example, the Democratic People's Republic of Korea). The countries that meet the requirements are listed in the appendix.

2.2 Notations

Abbreviation	Description
t	Time scale
l	Language
$P_{l' \to l}$	The shift possibility from language l' to language l
$Q_{l,t}$	The quantity of speakers of the language l in the year t
$T_{l,t}$	The number of global tourists who speak language l in year t
$tiev_{l,t}$	The total import and export volume assigned to language l
$\omega_{l,i}$	The proportion of the speakers of the language l in country i
G	The theoretical maximum global population growth rate
P_T	The total population of the world in the year t
K	The global carrying capacity
c	The language's maximum shift rate
$a_{l,t}$	Resistance level
$s_{l,t}$	Language status
D_l	The number of users joined the group about language l in Facebook
$R_{l,t}$	The proportion of countries with language l as their official language
V	The migration rate
M	The migration population

3 Model Construction

3.1 Worldwide Language Shift Model

We consider the influence of several factors and establish a language shift model to reveal the development principle of every language.

3.1.1 Factors

- Government Support

  Government support has a far-reaching impact on the development of language. The language established by the government as the official language can be used more widely, for example as an educational language among school teachers and students. Official languages tend to be more competitive and less likely to be eliminated than other non-official languages.

- Tourism

  With the improvement of people's living standards, the international tourism industry has also become increasingly prosperous. Tourism promotes the exchange of people of different languages. When tourists are traveling or living abroad for a short period of time, they will have an impact on the local language and culture, thus facilitating exchanges and collisions among different languages.

- International Business Relations

  With the development of globalization, international business trade is becoming more and more common. The economic exchanges between countries have led to the exchange of personnel and material resources, thus promoting the exchange and integration of different languages.

- Technology

  The progress of global science and technology, especially the development of online social media and the advancement of translation tools, have made communication and exchange between languages more frequent and more convenient. People can easily access various languages through the Internet to learn and exchange. As the translation software is more intelligent, people in different languages in different countries can understand each other.

3.1.2 Data Pre-processing

- Quantity of language speakers

  We have collected the worldwide number of language speakers of the selected 26 languages.Each of these languages has the number of native speakers and second or third language users, respectively. Due to limited data sources, we can only collect the data for the four years from 2014 to 2017 and use it as a basis for analysis and solution.We observe the changes of the number of native speakers and the total number of speakers, so as to observe the law of language development.

- Rate of being the official language

  We count the number of countries where each language is the official language and its ratio to the total number of countries in the world is the ratio of the language as the official language in 2017. In this way, we can measure the extent to which the language is supported by the government.In view of the fact that the official languages in each countries vary very little, here the proportion of each language as the official language is not subject to change over time. The proportion of countries with language l as their official language can be calculated as:

  $$R_{l,t} = \frac{N_o}{N} \tag{6}$$

  N_o is the number of countries with language l as its official language, N is the total number of countries. $N = 197$.

- Tourism Index

  We have compiled statistics on the number of departing population from each country in 2014-2017. We believe that the proportion of language users in each country's departing population is consistent with the overall rate of language speakers in the whole country, so that the number of languages spoken and outgoing by each country can be calculated. The calculation method is:

$$T_{l,t} = \sum_{i=1}^{N} T_{i,t} \times \omega_{i,l} \tag{7}$$

$T_{l,t}$ is the number of tourists all over the world who speak the language l and this number changes with time t. $\omega_{i,l}$ is the proportion of the speakers of the language l in nation i. As $\omega_{i,l}$ changes very small in recent years, we approximate it with its 2015 data.

- Total import and export volume

  We have collected the import and export trade data (in US dollar) of selected 36 countries in 2014-2016. We use the sum of the absolute values of imports and exports as the country's import and export indicators. We believe that the proportion amount of language shift brought about by a country's international trade with other countries is consistent with that of the national language. The total import and export volume assigned to language l speakers can be calculated as:

$$tiev_{l,t} = \sum_{i=1}^{N} tiev_{i,t} \times \omega_{i,l} \tag{8}$$

- Social Media

  When measuring the impact of social media on language shifting, we select the influential global social media Facebook as the data source. With the "group" lookup feature on Facebook, we look up the number of groups whose keywords are a language study group and count the number of users who participated in the group. For example, we enter the keyword "English study" in the group query interface, then we find out that there are 99 groups that contain this keyword, and the total number of participants in this kind of group is 65,109. Counting the number of language learning groups per language measures the contribution of social media (Facebook) to that language.

3.1.3 Worldwide Language Shift Model Construction

We expand the language competition model to a global scale, join multiple languages (not just a bilingual situation), and use the number of speakers of the language as a measure of language development. By exploring the influence of those factors mentioned above on the number of language speakers, we can predict the future development of language.This is Worldwide Language Shift Model:

$$\frac{dQ_{l,t}}{dt} = \sum_{l' \neq l} Q_{l',t} P_{l' \to l} - \sum_{l' \neq l} Q_{l,t} P_{l \to l'} + G Q_{l,t} (1 - P_t / K) \tag{9}$$

In this model:

- $P_{l' \to l}$ indicates the shift possibility from language l' to language l

- $Q_{l,t}$ is the quantity of speakers of the language l all around the world in the year t.
- G is the theoretical maximum global population growth rate.
- P_t is the total population of the world in the year t
- K is global carrying capacity. Due to the limited environmental resources, the maximum population that the world can accommodate is limited. According to Brown and Kane(1994), Daily and Ehrlich (1994, 1996), the global carrying capacity is 11 billion, so we assign $K = 1.110^10$.

From this expression we can see that for a language, its change over time can be calculated as $\frac{dQ_{l,t}}{dt}$.

$\sum_{l' \neq l} Q_{l',t} P_{l' \to l}$ illustrates the sum of the expected value of any language converted into the language l. This will lead to an increase in the number of language l speakers.

$\sum_{l' \neq l} Q_{l,t} P_{l \to l'}$ illustrates the sum of the expected value language l converted into any other language. This will lead to a decrease in the number of language l speakers.

$GQ_{l,t}(1 - P_t/K)$ illustrates changes in the number of speakers brought by the growth of the natural population.

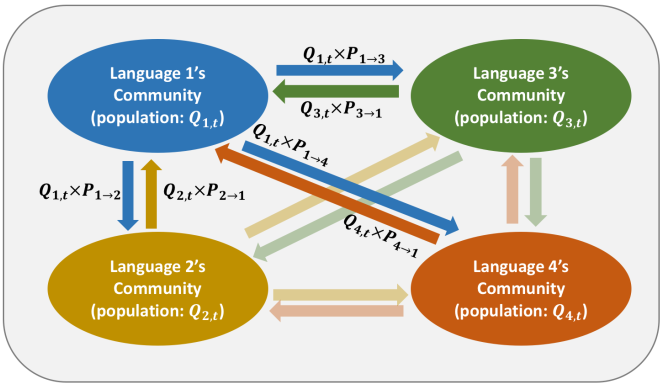


Figure 1: Global language shift schematic diagram

Shift Possibility

The shift possibility can be expressed as:

$$P_{l' \to l} = c \times s_{l,t} \times (Q_{l,t}/P_t)^{a_{l,t}} \tag{10}$$

In this expression:

- c is the maximum shift rate, the same as the A-S model, showing the maximum possibility of an individual to change his language.
- $a_{l,t}$ is the resistance level.$a_{l,t}$ quantifies the resistance level of other language speakers to change their language to language l. Resistance is the unwillingness to change to, learn and speak language l.

- $s_{l,t}$ is the language status, which indicates the social and economic opportunity and convenience of the language l. Being official language increases the social and economic opportunities and thus improves the language status.

Resistance Level

The resistance level can be expressed as:

$$a_{l,t} = -\frac{T_{l,t}}{P_t} \times \varphi - \frac{tiev_{l,t}}{GDP_t} \times \rho - \psi t - D_l \delta + \overline{a} \tag{11}$$

In this expression:

- $T_{l,t}$ is the number of tourists all over the world who speak the language l and this number changes with time t
- $tiev_{l,t}$ is the total import and export volume assigned to language l speakers
- ψ is the technology level increasing constant.
- D_l is the number of users joined the group about language l in the social media(statics from Facebook).
- $\overline{a}$ is a constant, the reference resistance level for each language. It indicates the difficulties encountered in learning and using a language, the resistance caused by social pressure and the average reluctance of people to change languages.

Tourism, international business relationship with other countries (which can be indicated by total the import and export volume) reduces the resistance.

Technological advances also make linguistic transitions more convenient. It lowers the difficulty of translation and facilitate online communication. Thus, over time, the resistance has an intrinsic decreasing trend.

The number of participants of the topic of a language in the online social network shows the popularity of this language, which shows the willingness and decreases the resistance.

The technology also reduce the language resistance as it lowers the difficulty of translation and facilitate online communication. Thus, over time, the resistance has an intrinsic decreasing trend.

Language Status

Language Status can be expressed as:

$$s_{l,t} = R_{l,t} \times \varsigma + \overline{a} \tag{12}$$

In this expression:

- $R_{l,t}$ is the proportion of countries with language l as their official language.
- $\overline{a}$ is a constant and is the baseline for each language status, indicating the average chance and convenience that one can get by using one language

To summarize, the unknown parameters in the above formula contain $G, c, \varphi, \rho, \psi, \overline{a}, \varsigma, \delta, \overline{s}$. Using the collected data, we can fit the parameters $\hat{G}, \hat{c}, \hat{\varphi}, \hat{\rho}, \hat{\psi}, \hat{\overline{a}}, \hat{\varsigma}, \hat{\delta}, \hat{\overline{s}}$ and obtain the development rules of various languages under the worldwide competition model.

3.2 Domestic Language Shift Model

Based on the established worldwide language shift model, we can extend the model to the domestic language competition. The principles and influencing factors are consistent with the global situation.

3.2.1 Domestic Language Shift Model Construction

For each country, we can also calculate the change of number of speakers over time as:

$$\frac{dQ_{i,l,t}}{dt} = \sum_{l' \neq l} Q_{i,l',t} P_{l' \to l} - \sum_{l' \neq l} Q_{i,l,t} P_{l \to l'} + \hat{G} Q_{i,l,t} (1 - P_{i,t}/K_i) \tag{13}$$

$$P_{l' \to l} = \hat{c} \times s_{i,l,t} \times (Q_{i,l,t}/P_{i,t})^{a_{i,l,t}} \tag{14}$$

$$a_{i,l,t} = -\frac{T_{i,l,t}}{P_{i,t}} \times \hat{\varphi} - \frac{tiev_{i,l,t}}{GDP_{i,t}} \times \hat{\rho} - \hat{\psi} t - D_l \hat{\delta} + \hat{\hat{a}} \tag{15}$$

$$s_{i,l,t} = R_{i,l,t} \times \hat{\varsigma} + \hat{\hat{a}} \tag{16}$$

The difference is that:

- K_i means the maximum population of each country. We assume the maximum is 1.5 times to the population in 2016. We will discuss about this assume later in the sensitivity analysis section.
- $R_{i,l,t}$ indicates that if the language l is the official language of country i.

$$R_{i,l,t} = \begin{cases} 1, l = o_i \\ 0, l \neq o_i \end{cases} \tag{17}$$

- $T_{i,l,t}/P_{i,t}$ is the proportion of the country's total tourist population in the country. We assume that it is a constant Γ calculated by the means of historical data:

$$\Gamma = \frac{1}{17} \times \sum_{t=2000}^{2016} \frac{T_{i,t}}{P_{i,t}} \tag{18}$$

- $tiev_{i,l,t}/GDP_{i,t}$ is the ratio of import and export volume to GDP in one country. We assume that it is a constant Ω calculated by the means of historical data:

$$\Omega = \frac{1}{17} \times \sum_{t=2000}^{2016} \frac{tiev_{i,t}}{GDP_{i,t}} \tag{19}$$

We get the new expression of resistance level:

$$a_{i,l,t} = -\Gamma \times \frac{Q_{i,l,t}}{P_{i,t}} \times \hat{\varphi} - \Omega \times \frac{Q_{i,l,t}}{P_{i,t}} \times \hat{\rho} - \hat{\psi} t + \hat{\hat{a}} \tag{20}$$

3.3 Migration Model

Migration is an important means of population mobility among nations. Studying the global migration trend is very important to solve the problem of language shift.

3.3.1 Data Pre-processing

We get the migration data from 1991 to 2017 with an interval of 5 years from United Nations(2017). The migration rate can be calculated as:

$$V_{ij,k} = \frac{M_{ij,k}}{5P_{i,k}} \tag{21}$$

In this expression:

- k is the count of time interval, i.e., $k = 1$ indicates the interval of 1991-1995 and $k = 6$ indicates the interval of 2016-2017
- $V_{ij,k}$ is the migration rate from country i to j in the interval k
- $M_{ij,k}$ is the net migration population from nation i to another nation j in the interval k.
- $P_{i,k}$ is the average population of the nation i in the interval k. For example, $P_{i,1}$ is the average population of the nation i from 1991 to 1995.

Let $\overline{V}_{ij}$ be the average of $V_{ij,k}$. The emigration from the nation i to another nation j of the year t is

$$M_{ij,t} = \overline{V}_{ij} P_{i,t} \tag{22}$$

3.3.2 Migration Model Construction

We believe that the world's future trend of population movement keeps the same trend. The migration rate in the future is equal to the average migration rate from 1991 to 2017.

It is easy to forecast the distribution of speakers of the country i at the end of the year t as:

$$Q_{i,l,t+1} = Q_{i,l,t} - \sum_{j=1}^{N} Q_{ij,l,t} + \sum_{j=1}^{N} Q_{ji,l,t} + E_{i,l,t} \tag{23}$$

In this expression:

- $Q_{i,l,t}$ is the speaker quantities of language l at the beginning of the year t in the nation i
- $Q_{ij,l,t}$ is the migrant quantities of language l in the migration route $i \rightarrow j$ in the year t. The reversed language flow is $Q_{ji,l,t}$.
- $E_{i,l,t}$ is the increase of speakers who speak language l in nation i in the year t because of the natural population growth.

$\sum_{j=1}^{N} Q_{ij,l,t}$ and $\sum_{j=1}^{N} Q_{ji,l,t}$ are total outflows and inflows of language l. To simplify the model, we assume that $Q_{ij,l,t}$ and $E_{i,l,t}$ are based on the year-beginning data. They can be calculated as:

$$Q_{ij,l,t} = Q_{i,l,t} \times \overline{V}_{ij} \tag{24}$$

$$E_{i,l,t} = G_i \times Q_{i,l,t} \times (1 - \frac{P_{i,t}}{K_i}) \tag{25}$$

where G_i is the maximum of intrinsic growth rate of the nation i, K_i is the population carrying capacity of the nation i.

To summarize, we get the expression about the speaker quantities as:

$$Q_{i,l,t+1} = Q_{i,l,t} \times (1 - \sum_{j=1, j\neq i}^{N} \overline{V}_{ij}) + \sum_{j=1, j\neq i}^{N} Q_{j,l,t} \times \overline{V}_{ji} + E_{i,l,t} \tag{26}$$

3.4 Geographic Language Shift Model

As globalization accelerates, cross-border language competition caused by migration is also more prevalent. We establish a geographic language shift model to explore the changes about various languages speakers across the globe.

3.4.1 Factors

- Migration

  The flow of population between countries will have an impact on the language conversion. Migration is an important factor affecting the language development model. Studying migration between countries can help us understand the changes in the number of people speaking different languages in different countries.

- Language Competition

  Migration does not directly determine the number of language speakers. When two or more languages are competing, the number of language speakers changes under the combined effect of multiple factors. This competitive model and the associated multiple factors have been discussed in the previous model.

3.4.2 Geographic Language Shift Model Construction

We combine the Domestic Language Shift Model and the Migration Model to explore the row of language shift between countries. The combined model is:

$$\Delta Q_{i,l,t} = \sum_{l' \neq l} Q_{i,l',t} P_{l' \to l} - \sum_{l' l/} Q_{i,l,t} P_{l \to l'} + G_i Q_{i,l,t} (1 - P_{i,t}/K_i) - Q_{i,l,t} \sum_{j=1, j\neq i}^{N} \overline{V}_{ij} + \sum_{j=1, j\neq i}^{N} Q_{j,l,t} \overline{V}_{ji} \tag{27}$$

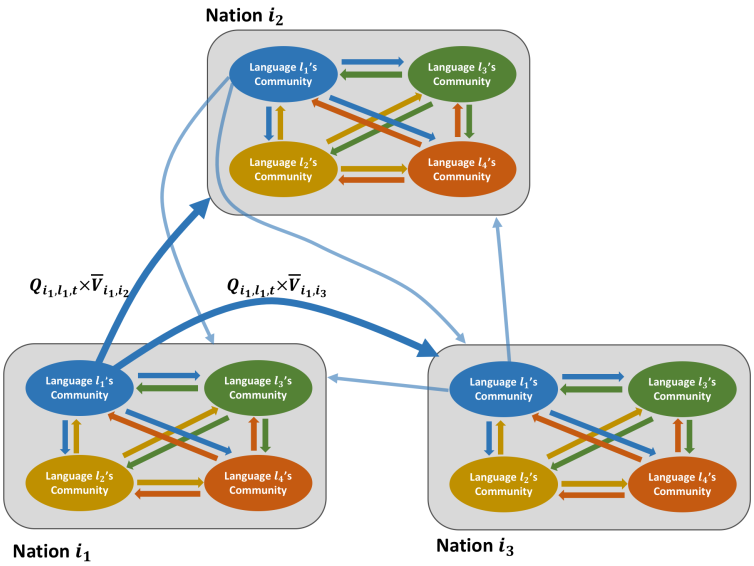


Figure 2: Geographical language shift schematic diagram

4 Model Implementation and Results

4.1 Part I Problem A

We model the distribution of various language speakers over time as the ***Worldwide Language Shift Model***.

For unknown parameters in the model, the gradient decent method is used to perform the fitting with programming tools.

Algorithm 1: Parameter Fitting

Step 1: Calculate the actual amount of change in the number of users for each language form 2014 to 2017

Step 2: Use the annual data in the model formula to calculate the predicted value of the change in the number of users for each language from 2014 to 2017

Step 3: The cost function is constructed as the sum of squares for error of the actual variation and the predicted quantity

Step 4: Solve the cost function to reach the minimum value of the parameters, that is, the result of the fitting

The result of parameter fitting is:

Table 1: Fitting Result

$\hat{G} = 1.838 \times 10^{-2}$	$\hat{\rho} = 1.246 \times 10^{-1}$	$\hat{\varsigma} = 1.392 \times 10^{-1}$
$\hat{c} = 2.43 \times 10^{-3}$	$\hat{\psi} = 2.093 \times 10^{-5}$	$\hat{\delta} = 2.128 \times 10^{-8}$
$\hat{\phi} = 1.054 \times 10^{-1}$	$\hat{\overline{a}} = 5.44 \times 10^{-1}$	$\hat{\overline{s}} = 4.962 \times 10^{-1}$

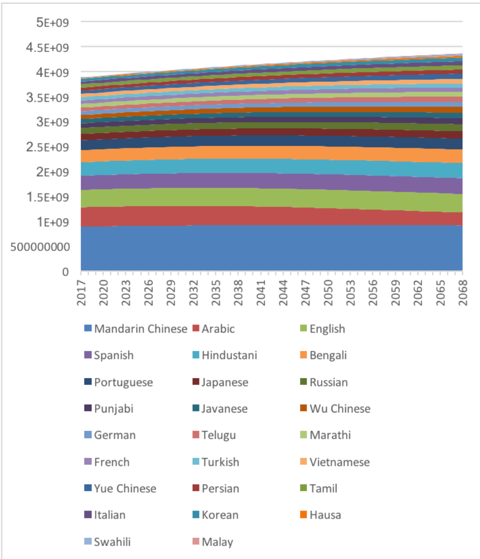
4.2 Part I Problem B

Use the ***Worldwide Language Shift Model*** to predict the future of speakers quantity in the next 50 years. We use the pattern:

$$P_t = P_{t-1}[1 + \hat{G}(1 - P_t/K)] \tag{28}$$

as the low of population growth. In fact, the choice of population growth mode has no essential effect on the result of our result. If the pattern of population growth changes, then we can calculate the development of the language under the new growth pattern.

Respectively, Use data of native speakers quantities of various languages and data of total speakers quantities of various languages in the model and solve the result for 50 years later. The changes of two indicators can be observed. Figure 3 and figure 4 indicates the number changes and Figure 5 and figure 6 indicates the proportion changes.



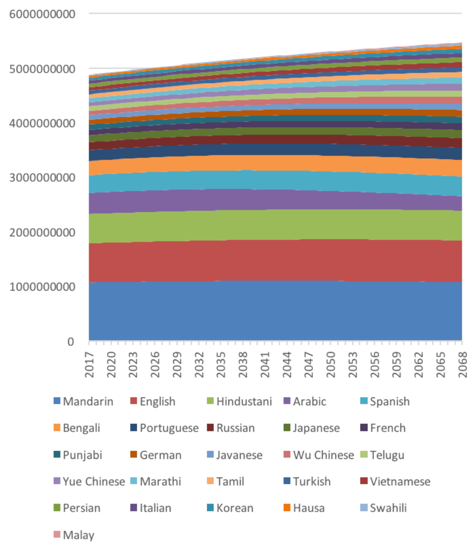


Figure 3: Native Speakers Quantities in 2068 Figure 4: Total Speakers Quantities in 2068

To see how rank of the top 10 languages changes, we plot the top 15 rank of languages according to native speakers quantities and total speakers quantities in 50 years. (Figure 7 and figure 8) Among the top 10 languages native speakers languages, only Punjabi (10th th) is replaced by Wu Chinese. Similarly, among the top ten total speakers languages, only French (10th) is replaced by Wu Chinese. This can be roughly attributed to the large number of tourists and foreign trade scale of people who speak Wu language. In addition, there are several internal adjustments in the order of the top ten in both rankings.

4.3 Part I Problem C

Use the ***Geographic Language Shift Model*** to predict the numbers of speakers of each language in each country for the next 50 years. In order to observe whether the geographical distribution of language has changed, one-way ANOVA was used.

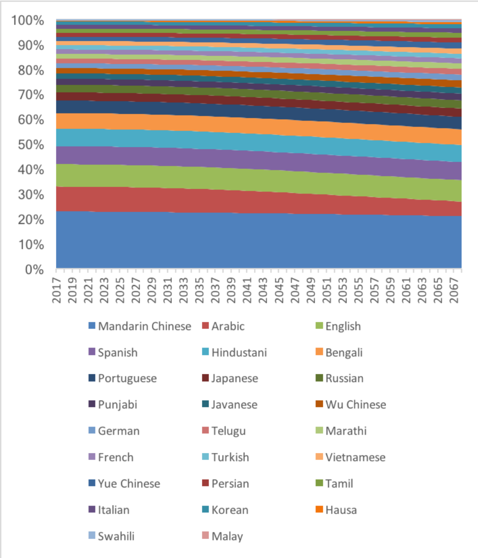


Figure 5: Native Speakers Proportions in 2068

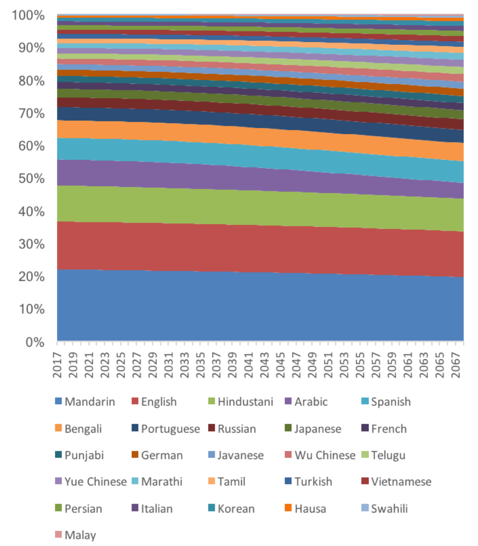


Figure 6: Total Speakers Proportions in 2068

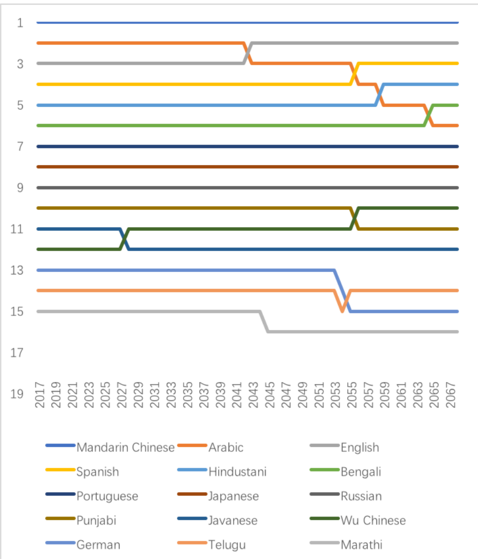


Figure 7: Native Speakers Rank in 50 years

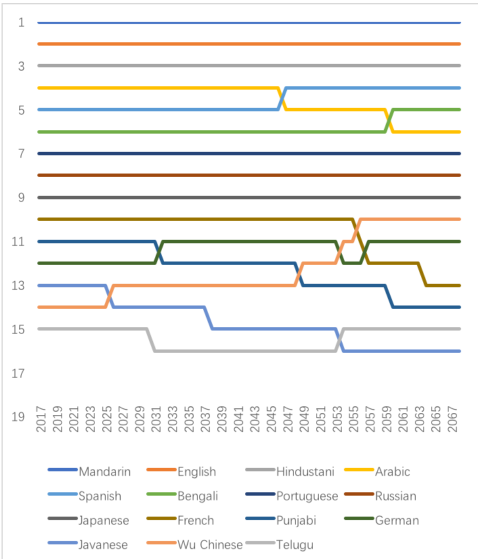


Figure 8: Total Speakers Rank in 50 years

Use ANOVA to test the number of users for each language in 2018 and the predicted number of users for each language in 2068. The result of ANOVA is listed in the appendix. According to the one-way ANOVA test, all languages pass the test except Malay. (This situation is because Malay is a union of several languages and its data is not very accurate.) In this way, we can say that other 25 languages geographical distribution will not change a lot in the next 50 years.

Use ANOVA to test the number of users for each country in 2018 and 2068 and the conclusions are that the geographical distribution in each country will not change during the next 50 years. The result of ANOVA is listed in the appendix.

We can observe the geographical distribution of English straightly from Figure 9 and 10.

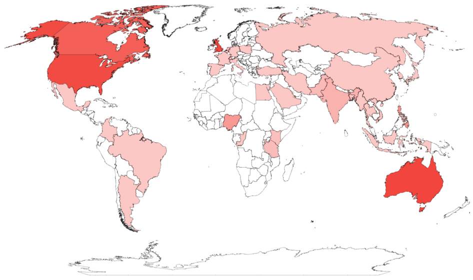


Figure 9: Distribution of English in 2018

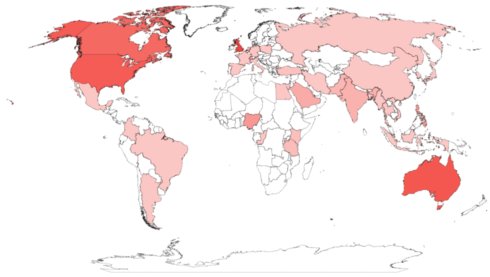


Figure 10: Distribution of English in 2068

4.4 Part II Problem A

We assume that each employee can speak multiple languages. The company needs to pay higher salaries to multilingual employees and such employees are harder to hire. We suppose that the company will pay C when each employee learn a new language. The benefit company can get by learning a new language l in one year is $Q_{i,l,t}/P_{i,t} \times F$. $Q_{i,l,t}/P_{i,t}$ indicates the business opportunity brought by learning a new language in one year and F indicates the business benefit one can get by earning one unit of business opportunity.

To get the maximum benefit, when recruiting staff, the company will give priority to employees who can speak both English and the language with the highest proportion in the local area.The net benefits each employee can bring to the company are:

$$\Lambda = Q(n)/P_{i,t} \times F - C \times n \tag{29}$$

$Q(n)$ is the accumulated number of speakers of English and the top $n-1$ languages in the local area. The expression $Q(n)/P_{i,t}$ illustrates the business opportunity one can get by learning n languages. The Λ will reach the maximum when $d\Lambda/dn = 0$, which means:

$$\frac{Q(n) - Q(n-1)}{P_{i,t}} = \frac{C}{F} \tag{30}$$

We can get the best employees number n^*

We assume that $C/F = 10\%$ and calculate n^* for each country. If $n^* < 2$ or English is not included in the top n^* languages, recruiting in that country will be ***uneconomical***.

Except for uneconomical countries, we rank the cities in other countries by the rank of ***City classification for 2016*** powered by *Globalization and World Cities Research Network* to choose the location of new offices. The result is:

We do not recommend recruiting employees with different language policies in the short term versus the long term. Because from the previous analysis, we know that the language distribution of these cities will not change a lot in the next 50 years.

Table 2: Office Locations

City	Country	n*	Corresponding language
Paris	France	3	France, Spanish, English
Tokyo	Japan	2	Japanese, English
Milan	Italy	3	Italian, France, English
Toronto	Canada	2	English, French
Bombay	India	3	English, Hindustani, Bengali
Brussels	Belgium	2	France, Dutch,English

4.5 Part II Problem B

As technology advances over time, cross-lingual communication becomes more and more convenient. This will reduce the advantages of being a particular language speaker and thus reduce the business opportunities and business benefits. On the other hand, the increase in the cost of employment will lead to an increase in C, so it will increase the ratio C/F over time. Based on previous studies, the proportion of the population for a language is basically constant. So n^* may decrease over time, thus failing to meet the company's requirements and rendering the office uneconomical. So in the long run, it may not be economical to have some offices whose $n^* = 2$.

In order to study this issue, we need exact C and F data, especially the salary data for employees who can speak different languages and the benefit data they bring to the company. In addition, there is a need for data that measures the level of technology, business communication efficiency and communication costs. Through these data, we can analyze change of cost C and return F under the influence of time and communication efficiency. In this way, we can get n^* to help the company make better decisions in the long run.

5 Sensitivity Analysis

5.1 Sensitivity Analysis for K_i

We have mentioned above that K_i is assumed as the 1.5 times of present population of nation i. In this part, we change the value of K_i to see whether it will significantly affect the global distribution of the population in all languages 50 years from now. Let the times $q = 1.25, 1.75, 2$ and do the ANOVA analysis on the result in 2068. The results are listed in the appendix.

According to the ANOVA analysis, both the number and the proportion of speakers are not sensitive to the change of K_i.

5.2 Sensitivity Analysis for domestic language distribution

Due to the error when collecting the data about Malay, we can not get the correct prediction result about Malay. As a result, we artificially add errors to the initial data to test the stability of the model. Using a random number generator, randomly select a country

i to change the initial language distribution of the country. Specifically, swap the number of users in the i^{th} language of the country with the number of users in the $L - i^{th}$ language. Due to limited computational power of the computer, 20 tests are conducted. Do the ANOVA analysis on the results of experiments. The ANOVA results are listed in the appendix.

The results show that there is no significant difference in the results of the model after all the data in one country are disrupted. That is to say, the distribution of each language is not affected by a single country. Data errors in Malay will not affect the results of the whole model.

5.3 Sensitivity Analysis for Fitting Parameters

$\hat{\hat{a}}$ is the based resistance level. $\hat{\hat{s}}$ is the based language status. $\hat{c}$ is the maximum shift rate. These three parameters impact the whole model during prediction and are very important. Use Austrilia as an example to explore the impact on the language distribution brought by the change of these three parameters. The results are showed in the Figure 11.

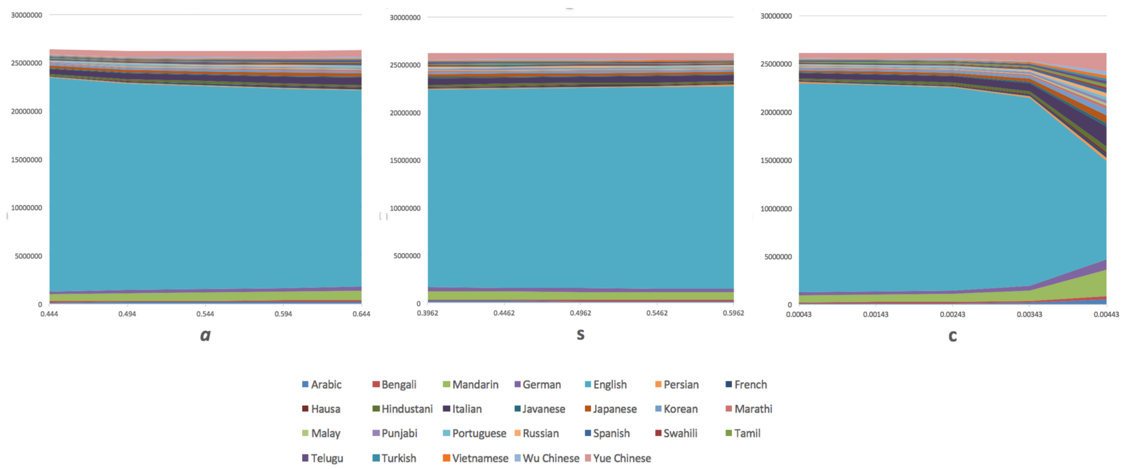


Figure 11: Sensitivity Analysis for Fitting Parameters

From the analysis, we can know that except for the positive increasing of c, the model is not sensitive to the change of a, s and the minus change of c. The model is very sensitive to the positive increasing of c. When the c gets lower, the main language will keep increasing because of its large base number. However, this enhanced trend will also begin to decline after a long enough period of time.

6 Strengths and weaknesses

6.1 Strengths

We pioneer and formulate the multi-language competition model and combine it with the global migration model, which is more practical than the bilingual competition model.

This composite model has a good interpretability and applies to language shifts both in the world and in a nation.

This model runs spontaneously after it is established and can predict the number of users of any language in any country at any one time. This model can smooth the impact of data errors caused by rough statistics.

We also set up another model that measures the economic marginal benefits from the perspective of the cost and profitability form the use of language, which is an innovation.

6.2 Weaknesses

At this stage, as there is limited access to linguistic data, we have simplified the model to some extent, for example, assuming the same maximum transfer speed between languages.

Before using the gradient descent method to find the optimal combination of parameters, a large amount of manual derivations and calculations are required. Moreover, this model has many parameters to fit, so the model's fitting time is long.

6.3 Further work

Since there still exists some weaknesses in our current work, we will continue to improve and optimize the model and solving process.

Firstly, we will look for more linguistic data and earlier data for each country to support the fit of the model and restore the simplified part of the model.

Secondly, we will discuss whether the parameters can be reduced with the assurance of accuracy. This can enhance the practical value of the model and reduce the data requirements.

Thirdly, we will use the statistical data to evaluate the forecasting results and improve the model once the statistics are released next time.

7 Conclusions

For Part I, we establish a language transfer model that integrates language competition and migration. We look for language distributions across countries and around the world for model fitting. After getting the fitted parameters, we predict the linguistic distribution of any group in the future, whether the group is the entire global, a country or a community. After inserting the initial distribution of each country's languages into the model, the model can automatically give the number of speakers each year, each country and each language through the method of Cellular Automata. After that, we predict that in the next 50 years, the distribution of various languages in the world will be basically stable.

For Part II, we set up another model, which gets the most economical hiring strategy based on the country's linguistic distribution. Then, we use this model to evaluate whether the world's major cities are suitable for the company to set up offices, and recommend six location options. We give advice based on the model, as communication becomes more convenient in the future, less than 6 offices should be established.

We also perform some further studies. We prove the stability and error tolerance of our model at the end of our discussion.

References

[1] Daniel M Abrams and Steven H Strogatz. Linguistics: Modelling the dynamics of language death. *Nature*, 424(6951):900, 2003.

[2] Lester Russell Brown and Hal Kane. *Full house: Reassessing the earth's population carrying capacity*. Earthscan, 1995.

[3] Gretchen C Daily and Paul R Ehrlich. Population, sustainability, and earth's carrying capacity. In *Ecosystem Management*, pages 435–450. Springer, 1994.

[4] Gretchen C Daily and Paul R Ehrlich. Socioeconomic equity, sustainability, and earth's carrying capacity. *Ecological Applications*, 6(4):991–1001, 1996.

[5] Courtenay Honeycutt and Daniel Cunliffe. The use of the welsh language on facebook: An initial investigation. *Information, Communication & Society*, 13(2):226–248, 2010.

[6] Anne Kandler, Roman Unger, and James Steele. Language shift, bilingualism and the future of britain's celtic languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559):3855–3864, 2010.

[7] April MS McMahon. *Understanding language change*. Cambridge University Press, 1994.

[8] Salikoko S Mufwene. *The ecology of language evolution*. Cambridge University Press, 2001.

[9] Katharina Prochazka and Gero Vogl. Quantifying the driving factors for language shift in a bilingual region. *Proceedings of the National Academy of Sciences*, 114(17):4365–4369, 2017.

[10] Department of Economic United Nations and Population Division (2017) Social Affairs. Trends in international migrant stock: The 2017 revision (united nations database, pop/db/mig/stock/rev.2017). `http://www.un.org/en/development/desa/population/migration/data/estimates2/estimates17.shtml`.

[11] Ethnologue `https://www.ethnologue.com/statistics/size`

[12] Wikipedia `https://en.wikipedia.org/wiki/Main_Page`

[13] The World Bank `https://data.worldbank.org`

Appendices

Table A1: Country List

China	Egypt	Netherlands	Tanzania
Argentina	France	Nigeria	Thailand
Australia	Germany	Pakistan	Turkey
Bangladesh	India	Philippines	United Kingdom
Belgium	Indonesia	Poland	United States
Brazil	Iran	Russian Federation	Viet Nam
Myanmar	Italy	Saudi Arabia	
Canada	Japan	South Korea	
Colombia	Kenya	Spain	
Congo	Mexico	Switzerland	

Table A2: Language List

Mandarin Chinese	Hausa	Tamil
English	Punjabi	Marathi
Hindustani	German	Yue Chinese
Spanish	Japanese	Turkish
Arabic	Persian	Vietnamese
Malay	Swahili	Italian
Russian	Telugu	
Bengali	Javanese	
Portuguese	Wu Chinese	
French	Korean	

Table A3: Language ANOVA

Language	**F-value**	**P-value**
Arabic	0.126460763	0.723199651
Bengali	0.01946812	0.889433442
Mandarin	1.57292E-06	0.999002891
German	0.081672152	0.775888019
English	0.009311393	0.923402739
Persian	0.075428564	0.784399768
French	0.055489296	0.814460637
Hausa	0.225852825	0.636095674
Hindustani	0.000111288	0.991613026
Italian	0.060315915	0.806715958
Javanese	0.054260691	0.816488366
Japanese	0.015521312	0.901209534
Korean	0.17026054	0.681141448
Marathi	0.088763071	0.76663808
Malay	7.660217524	0.007217351
Punjabi	0.044877544	0.832844964
Portuguese	0.008953098	0.924886351
Russian	0.02091572	0.885424378
Spanish	0.03310032	0.856159619
Swahili	0.610550371	0.437215462
Tamil	0.103668634	0.748430346
Telugu	0.081301552	0.776383223
Turkish	0.02908759	0.865068954
Vietnamese	0.08705579	0.768827363
Wu Chinese	0.114672547	0.735899623
Yue Chinese	0.142967776	0.706492498

Table A4: Country ANOVA

Country	F-value	P-value
Argentina	0.007180598	0.932807751
Australia	0.010279037	0.919649862
Bangladesh	0.005194494	0.942831475
Belgium	0.036227065	0.849818022
Brazil	0.007068465	0.933333186
Canada	0.01255359	0.91123807
China	0.009238495	0.923811796
Colombia	0.00719267	0.932751431
Congo	0.008827832	0.925519183
Egypt	0.005450956	0.941439765
France	0.02490977	0.875227783
Germany	0.006832509	0.934452725
India	0.030554174	0.861944636
Indonesia	0.00934412	0.923378874
Iran	0.007673707	0.930544745
Italy	0.009250409	0.923762838
Japan	0.007241333	0.932524882
Kenya	0.009448448	0.922953684
Mexico	0.007516689	0.931257174
Myanmar	0.133593491	0.716276655
Netherlands	0.369745359	0.545895404
Nigeria	0.01762093	0.894928753
Pakistan	0.016767358	0.897490374
Philippines	0.01076275	0.917787779
Poland	0.000817746	0.977300434
Russian Federation	0.005671598	0.940268571
Saudi Arabia	0.000110737	0.991645751
South Korea	0.012061057	0.912989485
Spain	0.007454537	0.931541246
Switzerland	0.00700541	0.933630498
Tanzania	0.013530812	0.907863293
Thailand	0.009630563	0.922217114
Turkey	0.063482105	0.802107443
United Kingdom	0.00751444	0.931267433
United States	0.01171465	0.914243064
Viet Nam	0.00867058	0.926183563

Table A5: ANOVA after Distribution Modified

No.	F-value	P-value
1	0.012109461	0.912690158
2	2.09116E-06	0.998850306
3	0.054184749	0.816614494
4	0.005367356	0.941806384
5	0.051066312	0.821876616
6	0.03653546	0.848967267
7	4.8344E-05	0.994472145
8	0.040907493	0.840303268
9	0.03702232	0.847976735
10	0.010132928	0.920106186
11	0.059291545	0.808331394
12	0.029920096	0.863170845
13	0.005664064	0.940222538
14	0.013522629	0.907758193
15	0.020080842	0.887718584
16	0.069403299	0.792981514
17	0.054251554	0.816503536
18	0.019573792	0.889135748
19	0.06004562	0.807140792
20	0.075328546	0.784539154

Table A6: ANOVA Analysis of the Language Populations in 2068

q	F-value	F-crit
1.25	0.6184	4.0343
1.75	0.7477	4.0343
2	0.9201	4.0343

Table A7: ANOVA Analysis of the Language Population Proportions in 2068

q	F-value	F-crit
1.25	8.8131×10^{-5}	4.0343
1.75	0.0177	4.0343
2	0.2402	4.0343

Table A8: City Classification for 2016

1	London(UK)	11	Frankfurt(Germany)
2	Singapore	12	Madrid(Spain)
3	Paris (France)	13	Warsaw(Johannesburg)
4	Tokyo (Japan)	14	Toronto(Canada)
5	Dubai (UAE)	15	Bombay(India)
6	Brussels (Belgium)	16	Seoul(Korea)
7	Milan (Italy)	17	Istanbul(Indonesia)
8	Chicago (Chicago)	18	Amsterdam(Netherlands)
9	Mexico City (Mexico)	19	Brussels (Belgium)
10	Moscow(Russia)		