For office use only	Team Control Number	For office use only
T1 ________________	89996	F1 ________________
T2 ________________		F2 ________________
T3 ________________	Problem Chosen	F3 ________________
T4 ________________	B	F4 ________________

2018
MCM/ICM
Summary Sheet

There are a total of around seven thousand languages spoken on Earth on a daily basis. From the prevalent languages such as Chinese and English to the anonymous dialect of Hindi somewhere deep in the mountains of India, the total number of speakers and their geographic locations consistently changes over time. This change may be due to a number of factors: the newly-born taking their first trial of native language, the immigrants learning new languages to fit in the foreign environment, or students studying a second tongue according to the requirement set by the schools . . . This paper has explored such changes and built various models to predict the future trend of speaker both in amounts and geographically.

The first set of models is devoted to the studying of native and total speaker amount variation. By first making a bold and innovative connection between the similarities of languages and diseases, the discrete SIR model is applied with corresponding changes to predict the trend of total speakers. Basing upon this model and its principles, improved models are built successively to give analytical expressions with differential equations, incorporate multi-lingual interactions by Euler's Method, and separate influencing factors such as immigration and tourism using matrices. Each fit for different situations, these models have accurately predicted the trend of total and native speakers in the future 50 years. As the results show, English would have become the most popular language, while Javanese drops out of Top 10 and is replaced by French at No. 8.

As for the geographical distribution, this paper first developed models on global populations (the logistic model) and human migration patterns (gravity model of migration), and then using these models to consider the speaker composition of a variety of places by programming. The exact steps of programming and interaction of parameters have been explained in detail, and a graphic presentation of the language speaker distribution is hence presented using MATLAB.

Afterwards, paper goes on to investigate its application in reality by using the predicted data to determine the sites of new offices of a company with consideration of the local language composition and communication. Analyzing the situation from the perspective of the company, an evaluation criterion is designed combining all of language, economic, and location factors. Using the AHP model to weigh the important of each of these factors, a site selection model is hence built and exercised by Python to calculate the best options among the economics biggest cities in the world. Eventually, incorporating long term and short term expectations, a suggestion is made for the company to choose their new companies in cities such as Amsterdam and Paris (as an example). The diminishing marginal increase of the criterion also helps to determine strategies of opening less companies.

Overall, the model is filled with innovative ideas and practical applicability. The comparison between the prevailing of languages and diseases is not only creative but factually based, and conclusion drawn from which has become an essential idea in this paper. The decomposition of multiple influencing factors using transformation matrices and vectors is not only effective in its clean separation but also mathematically elegant. Plus, the graphic representation of result is direct and understandable by decision-making officers from the company. The short-long run consideration, along with the consistent result from sensitivity analysis, increases the adaptability of the model. The very high R^2 values throughout the regressions also indicate the practical values of this model.

In conclusion, this model incorporates considerations of various aspects and therefore has produced a model that is not only mathematically elegant but is also able to comprehensively make decisions for the company.

Key Terms: Native Speakers, Geographical Distribution, SIR Model, Logistics Model, Analytical Hierarchy Process (AHP), Transformation Matrix, Euler's Method, Gravity Model of Migration, K-Means Clustering.

Predicting the Distributions of Language Speakers and Its Application in Office Site Selections

1. Introduction and Problem Restatement

Among all of the 6900+ languages spoken on Earth, there are some that are prevalent with speakers everywhere while some anonymous or even at the edge of distinction. This model focuses on this problem, and predicts the future trends of popular languages including both the total numbers of their speakers and their geographical distributions. Having this prediction, the paper goes on to investigate its application in reality by using the predicted data to determine the sites of new offices of a company with consideration of the local language composition and communication.

Part I of the problem mainly requires us to do prediction from the language's side solely, both the trend of native and total speakers and their geographical distribution patterns. Part II of the problem concerns the background of the office site selection problem, and the company needs to consider this siting problem under a variety of conditions such as office number and time span.

2. Variables Assigned and Terminologies

Variable	Explanation	Units	Variable	Explanation	Units
N	Global Population	Millions	$I_{A\to B}$	Number of Immigrants from Language Zone A to B	Millions
n	Population of a country	Millions	BR_{AB}	# of Business Relations between Language Zone A and B	Thousands
N_i	Number of Native Speakers in a country in year i	Millions	$T_{A\to B}$	Number of Tourists from Language Zone A to Zone B	Thousands
b	Birth Rate	%	Te	Coefficients for level of technology	N/A
d	Decease Rate	%	E_A	Number of Students that Study Language A	Millions
t	Time	Year	$k_A\ or\ k_A{}'$	Constant	N/A
S or S_t	Number of Total Speakers of a language	Millions	C_A	Influence of pop culture	N/A
S_{1A}	Number of Native Speakers of Language A	Millions	P_{AB}	Usage of language A in Government B	True/False
S_{2A}	Number of Non-native Speakers of Language A	Millions	L_i	Languages	N/A
$P_{A1}, \dots$	Total Population in city A_1	Thousands	S_{A_1A},$S_{A_1B}\cdots$	Total Speaker of language A, B, … in city A_1	Thousands
$L_A, L_B, \dots$	Number of total speakers of language A, B, …	Millions	$A_1, A_2, \dots$	Cities in language A's language zone	N/A

Language Zone (of a certain language):

The combination of the places globally that uses this certain language as their main language.

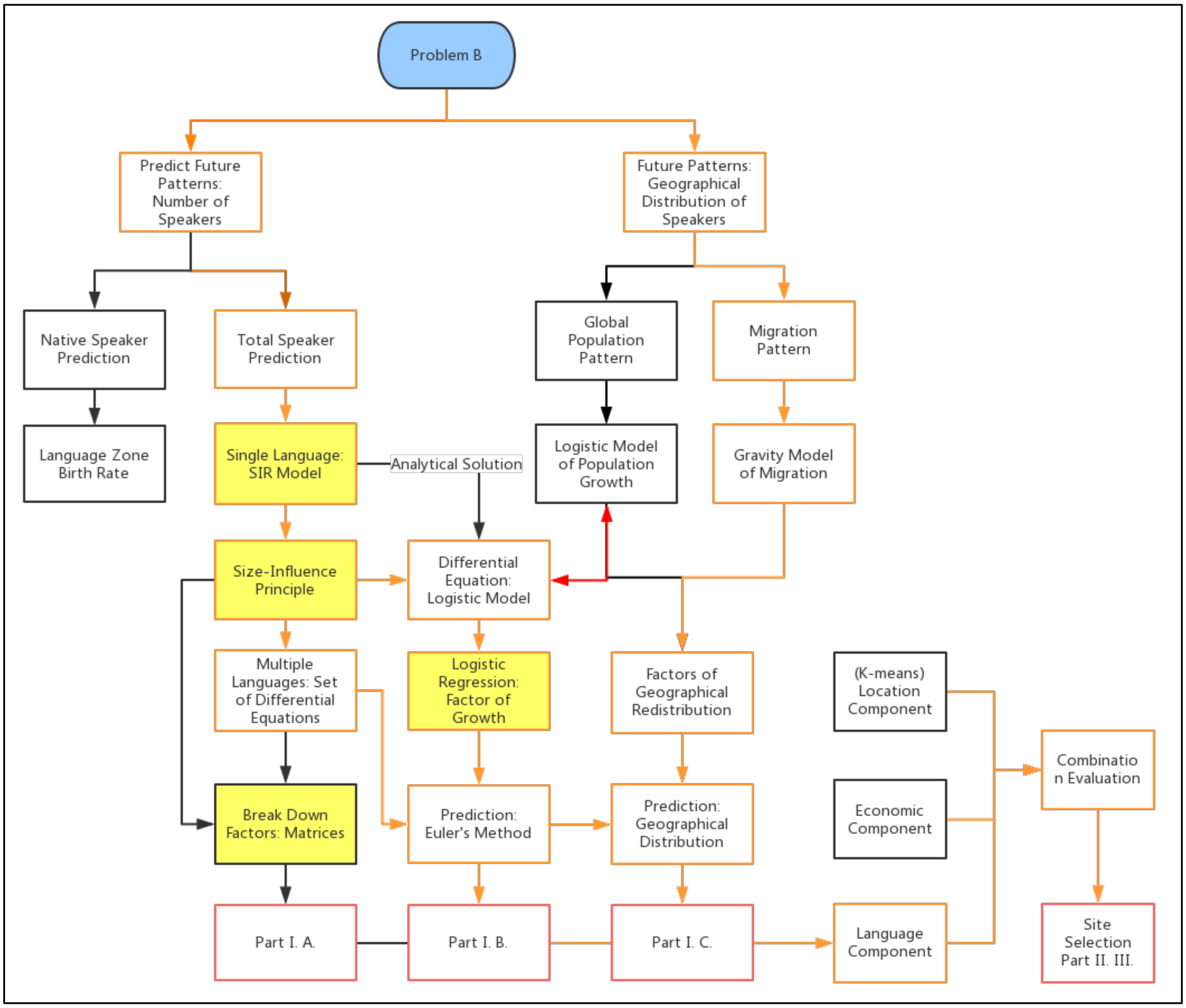
Size-Influence Principle (will be discussed again later):
The influence of a language, like diseases, is directly related to the number of speakers.

3. Assumptions and Justifications

1. People would forever remember a language once learned. Justification: In order to simplify calculations and limits ways how language speakers could diminish other than deceases.
2. There's no upper limit to the number of language a person would learn. Justification: Still for the sake of simplification to prevent the need to specify every person's language abilities.
3. People learn languages based on needs, ignoring the effect of linguistic interests. Justification: this assumption is made to minimize the categories, especially inconsistent sources of motives.
4. Immigrants move to places with higher living standards and income, not lower. Justification: In order to better predict the migration pattern; the main prediction model could also work otherwise but this is not the focus of this paper.

4. Model Construction

4.1. Overview of Methodology


Problem B
Predict Future Patterns: Number of Speakers
Future Patterns: Geographical Distribution of Speakers
Native Speaker Prediction
Total Speaker Prediction
Global Population Pattern
Migration Pattern
Language Zone Birth Rate
Single Language: SIR Model
Analytical Solution
Logistic Model of Population Growth
Gravity Model of Migration
Size-Influence Principle
Differential Equation: Logistic Model
Multiple Languages: Set of Differential Equations
Logistic Regression: Factor of Growth
Factors of Geographical Redistribution
(K-means) Location Component
Combinatio n Evaluation
Break Down Factors: Matrices
Prediction: Euler's Method
Prediction: Geographical Distribution
Economic Component
Part I. A.
Part I. B.
Part I. C.
Language Component
Site Selection Part II. III.

4.2. Language Speaker Variation Model

First of all, in order to eventually contribute to the selection of office site considering the factor of language, we need to investigate the global trend of the number of speakers for each of the languages, meeting the requirement of **A**, **B** in **Part I**. In this model, the variation of the amount of speakers for the top 20 languages over time is investigated step by step, starting from the simpler case of considering only one language to the more complex issue of the interactions of multiple languages at the same time. Building upon and inspired from the previous sub-models, each section represents a step in the thought process of creating these models.

4.2.1 is devoted to investigating the variations of the number of native speakers, while 4.2.2 and 4.2.3 are focusing on the total numbers of speakers. 4.2.4 combines the previous models and put them into numerical use, answering the questions directly by making corresponding predictions.

4.2.1. Native Speaker Variation

The total number of speakers could be divided into to two components: Native Speakers and Second (or 3rd, etc.) Language Speakers (Non-native Speakers). The following relation clearly holds:

$$S = S_1 + S_2$$

In this section, we mainly focus on the study of the variation of S_1.

Native Speakers are those "who has spoken the language in question from the earliest childhood." [1] Therefore, only those growing up in the Language Zone (defined in this paper as the combination of the countries that share a mainly used language) of a certain language are native speakers. In this paper, the term Language Zone would be the focused subject of discussion rather than countries or regions.

Since every Language Zone is made up of all of its countries, clearly $S_i = \sum N$, hence the task becomes calculating the native speakers in every country. The following relation holds:

$$N_i = (1 - d)N_{i-1} + bn_{i-1}$$

In the above equation relates the number of native speakers in a country, N_i ,with the number a year before, N_{i-1}. Given the death and birth rate of the country, the native speakers would be decreased by timing a factor of $(1 - d)$ due to the amount of the deceased. In comparison, recognizing that all newborns in this country are becoming native speakers of the language because of the environment, no matter if the parents are native speakers or just immigrants. Therefore, N_i is also increased by bn_{i-1}, the product of the birth rate and the total amount of citizens. By using this equation, one year's data of native speakers could be used to predict next year's, hence accomplishing the task of prediction.

4.2.2. Single Language Variation Sub-Model

A. SIR Model and the Size-Influence Principle

While considering the variation of total speakers of a certain language, the amount would become much more complex because it involves people's motives of learning a second language and becoming a Non-native Speaker. After considering this issue for a while, a creative connection was made.

Considering how one gains the motive to pick up a new language. The language must be very influential in the world, having many speakers, either native or non-native, such that learning this language could benefit them a lot. One could be able to communicate freely with many more people around the globe, collecting information from foreign sites written in this language, and so on. All of which are basing on the size of current Total Speakers, representing how influential and mainstream the language is.

We have realized that this would be very similar to the spreading of a disease; except for the difference in passively catching a disease and actively deciding to learn a new language, the strengths of influence are both basing on the size of the "infected." The larger the infected population, the more likely one is going to catch a disease; similarly, the larger the amount of Total Speakers, the more prompted one is to learn the language. We would later on refer to this principle inspired by SIR model the "**Size-Influence Principle**": the larger the size is, the larger its influence would be. Therefore, the discrete SIR model for disease spreading is applied here to show this inner connection, attempting to mathematically express the effect of the number of speakers on people's learning motives.

SIR model is short for "Susceptible Infective Removal Model," which is a typical model used in predicting the size of the infection. This fits our requirement exactly, because we are just looking for the size of the population "infected" by a certain language. The model is general in analyzing the influential factors, which would only be included in the parameters but not further broken down. More detailed modeling would base on its essential idea later on in 4.2.3.

The SIR model is usually consisted of three equations, describing the variation of the "susceptible" population, the "infective" population, and the "removal" population. Susceptible population corresponds to the population that could be potentially infected, and hence in this case, it is those who might have the need to learn the language for reasons such as business cooperation, assimilating immigrations, or education requirements. Infective population is the Total Speaker of the language, and one could only be moved out of the system in cases of deceases. The following set of relations is proposed over a time unit:

$$\begin{cases} S(n+1) = S(n) - aS(n)I(n) + f(n) \\ I(n+1) = I(n) - dI(n) + aS(n)I(n) + bn_i \\ R(n+1) = R(n) + dI(n) \end{cases}$$

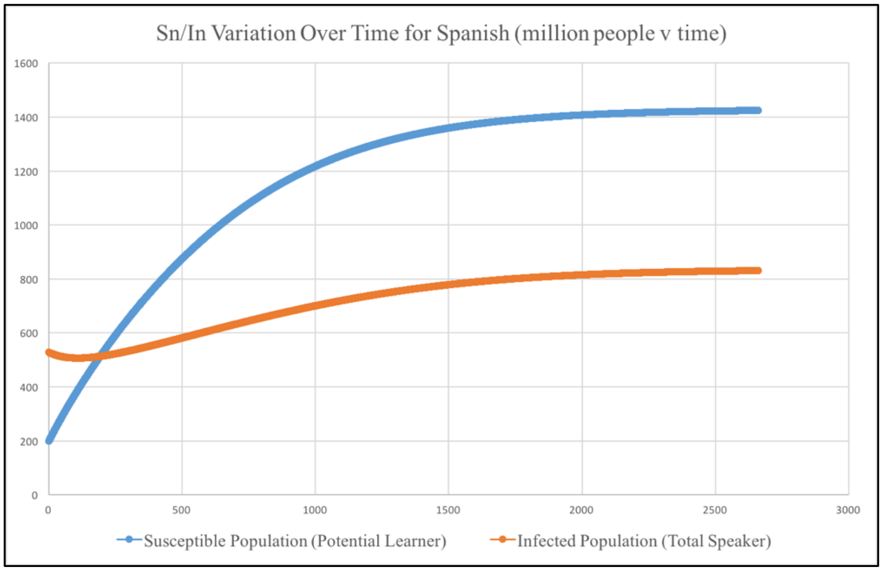
(The continuous case could be simply derived by substituting ΔS with $\frac{dS}{dt}$ and so on.)

In the first two equations, term $aI(n)$ is the rate at which susceptible people are urged to learn and become a Non-native speaker successfully, which is proportional to $I(n)$ because the Total Speaker population represents the language's global influence by our previous discussion of the **Size-Influence Principle**. Therefore, $aS(n)I(n)$ is the population moved out of S into I.

The rest of the terms in the second equation, $(I(n) - dI(n) + bn_i)$, is based on the Native Speaker variation equation, where Total Speaker is deceasing at a certain rate while new Native Speakers are born at a certain rate natively. It is noticeable that $f(n)$ represents the newly interested potential learners being added into the susceptible "pool," which is further broken down by the following expression:

$$f(n) = \frac{I(n)}{N} \times k$$

This is another application of the **Size-Influence Principle**. The newly susceptible population is directly related to the proportion of the Total Speakers in the world. Using Euler's Method in Excel, we plotted the following population variation considering only one language:



This gives a first-step general prediction. Clearly, both population for the language Spanish tends to stabilize over time, which means that the total number of speaker would eventually end up at around 830 million for Spanish, whose current data is at 528 million. After analysis of the original function, clearly, when

$$\begin{cases} f(n) - aS(n)I(n) = 0 \\ bn - aS(n)I(n) - dI(n) = 0 \end{cases}$$

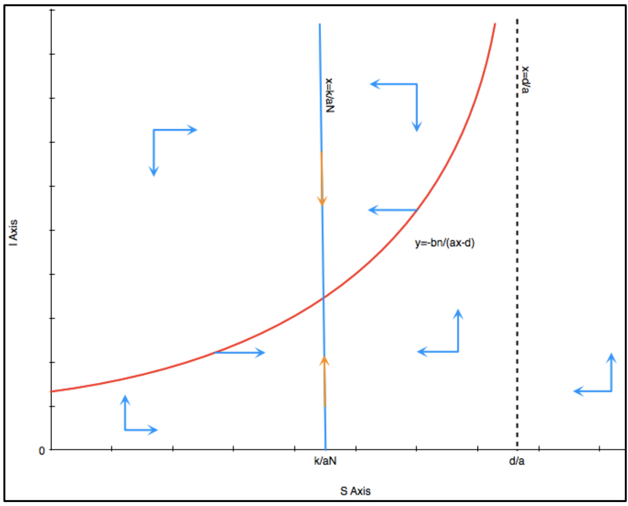
holds, both $I(n+1) = I(n)$, $S(n+1) = S(n)$. The solution gives the stabilized population size combination:

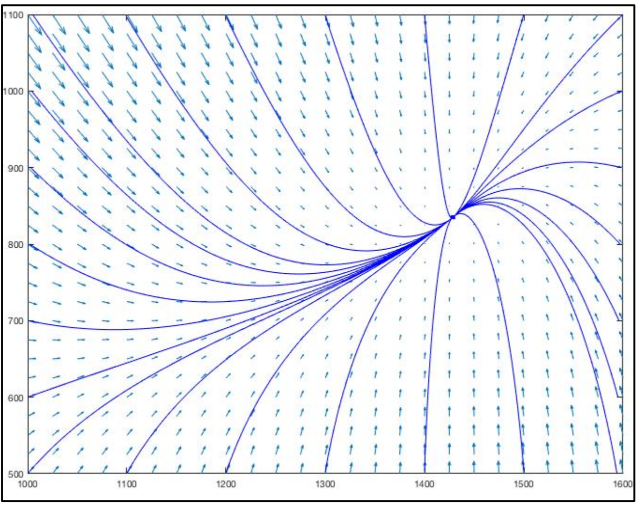
$$S(n) = \frac{k}{aN}, I(n) = \frac{Nbn_i}{Nd - k}$$

This does obey the common sense in that any change in the parameter does qualitatively follows what one would expect by the following table:

When	$I(n)$ or $S(n)$	Explanation
d increases	$I(n)$decreases	Higher death rate means less stabilized level of population
N increases	$S(n), I(n)$ decreases	Higher earth population means relatively less dense the population of this language, hence less language influence
b increases	$I(n)$ increases	Higher birth rate means higher stabilized level of population
n_i increases	$I(n)$ increases	More citizen means more stabilized Native Speakers

Another question to raise here is, does the model stabilize to this level no matter where the starting point is? By doing an analysis of the signs of ΔS $(S(n+1)-S(n))$ and ΔI $(I(n+1)-I(n))$, we plotted the graph showing the nature of the trend/movement of any point (S, I) on the Cartesian Plane (left), along with the Vector Field and Streamline plotted using MATLAB:





As one could clearly see from both of the diagrams, the intersection of the equations above is the point that all points would stabilize towards, hence an overall stable point is found, and the prediction is given by Euler's Method.

B. Differential Equation and Logistic Model

However, the weakness to the previous SIR model is that there is no analytical solution. The second-degreed differential equation derived from the SIR model with highest power of 2 is unsolvable, and hence a simpler model considering only the amount of Total Speakers is developed using only one single variable. The analytical solution is very important in terms of regressions to fit the current data as well as hence predicting future patterns, which explains the necessity of this model. Yet, this model is basing itself completely on the elements of the SIR disease model and applying the experience and knowledge learned previously to treasure this special connection made between language and diseases.

Allow me to introduce the following differential equation first modeling the change of a single language's Total Speaker change over time:

$$\frac{dS_t}{dt} = k_1 S_t (N - S_t) + (b - d) S_t$$

The first term $k_1 S_t$ represents the addition of new speakers: the growth by motives for people to pick it up as the second language is proportional to the size of the current Total Speaker population, or a.k.a. by the **Size-Influence Principle** concluded from the SIR model. According to the assumption that languages won't be forgotten once learned, $k_1$is positive.$(N - S_t)$ is the logistic term to control the size of the Total Speakers to not exceed all human population, making an analogy to the Population Model raised by Malthus in 1798. The next term, $(b-d)S_t$, is the natural birth rate indicating the change in native speakers deceased and born.

Solving the above equation,

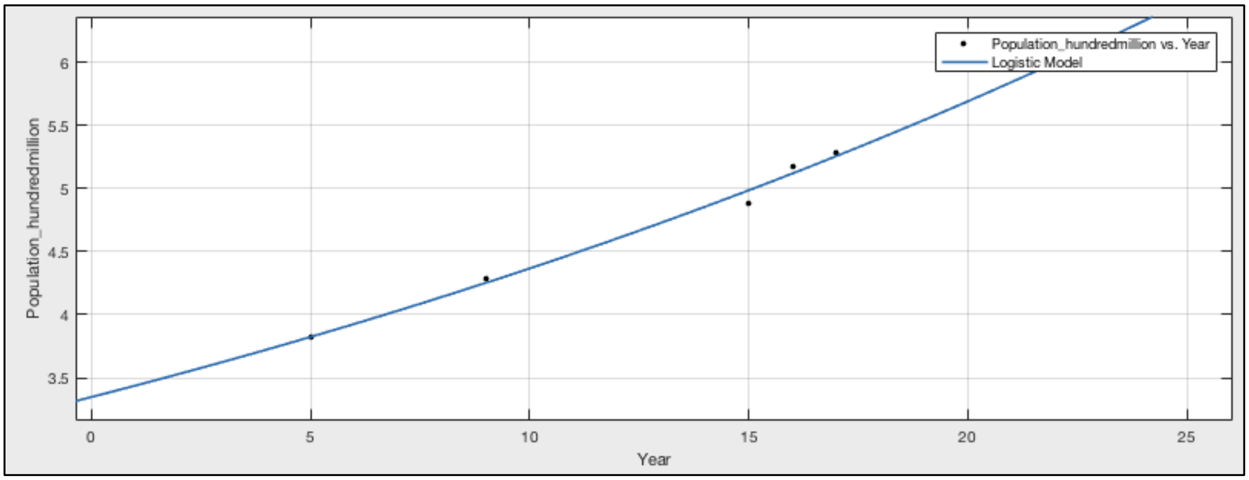
$$\frac{dS_t}{dt} = k_1 S_t\left(N + \frac{b-d}{k_1} - S_t\right)$$

$$\int \frac{dS_t}{S_t(N + \frac{b-d}{k_1} - S_t)} = \int k_1 dt$$

$$S_t = \frac{(Nk_1 + b - d)ce^{(Nk_1+b-d)t}}{k_1(1 + ce^{(Nk_1+b-d)t})}$$

Clearly, the result is a logistic model with maximum population $(N + (b-d)/k_1)$ and factor of growth $(Nk_1 + b - d)$, which is an instant reflection of how fast the size of Total Speaker is growing. Logistic models have two asymptotes which are the level of stabilization similar to $I(n)$'s in the SIR model. There are 2 constants yet to be determined from actual data: c and k_1.

Using the curve fitting tool in MATLAB, we are allowed to model upon actual data. Here is the example this paper shows here of the total speaker data for Spanish:



$$S_t = \frac{1115e^{0.027t}}{329.7 + 3.381e^{0.027t}},\ R^2 = 0.99$$

The very high R^2 value means that the equation is very well chosen, and the model fits reality quite perfectly. We would utilize this model again in 4.2.4, for calculations of predictions as one of the proposed measures and data supplier.

4.2.3. Multiple Languages Variation Sub-Model

C. Multiple Language General Model

Coming into considering multiple languages' interaction, the problem becomes complicated again. Still, we would consider the simpler case first by considering the interaction between two languages first, and then go into a multi-variable dynamic system.

Again, basing off of the results of our previous discrete SIR model and the Size-Influence Principle, we have proposed the following general model:

$$\begin{cases} \dfrac{da}{dt} = k_a \dfrac{a}{b}(N - a) \\ \dfrac{db}{dt} = k_b \dfrac{b}{a}(N - b) \end{cases}$$

In the above set of differential equations, a and b denote the amount of Total Speakers for language A and language B, respectively. Here, the model relies on the relative influence between the two languages. The greater a means the greater influence of language A according to the **Size-Influence Principle**. However, language B also has its own influence, which may cause those already speaking language B to be reluctant to acquire a new language seeing how themselves already speak a prominent language when b is large.

Consider the following scenario: the population of language A speakers of size 50 and language B speakers of size 100 are mixed and hence interacting with each other. Obviously, they would start learning each other's language. However, language A speakers would have more motive to learn than language B speakers, because they are the minority here and being able to communicate with a lot more people would be beneficial on a larger scale than that benefit to learn A for language B speakers. Their relative size is a representation of the relative influence upon one another, hence the ratio $\frac{a}{b}$ is used in the model.

The term $(N-a)$ and $(N-b)$ act as a logistic term to prevent from situations where the speakers of language exceeds, in size, the whole human population. k_a and k_b here indicates the coefficient of the motive, which might be dependent on a variety of factors such as the influence of the country, government policy, or the difficulty of the language. It directly affects the speed at which a language speaker population would grow, similar to the factor of growth in the logistic model. The further breakdown would not be considered here but in the following model B with matrices.

Further extending this model using similar arguments would give as the following set of equations:

$$\begin{cases} \dfrac{dL_1}{dt} = k_{L_1} L_1 (N - L_1) (\dfrac{1}{L_2} + \dfrac{1}{L_3} + \dfrac{1}{L_4} ... + \dfrac{1}{L_m}) \\ \dfrac{dL_2}{dt} = k_{L_2} L_2 (N - L_2) \left(\dfrac{1}{L_1} + \dfrac{1}{L_3} + \dfrac{1}{L_4} ... + \dfrac{1}{L_m}\right) \\ \dfrac{dL_3}{dt} = k_{L_3} L_3 (N - L_3) \left(\dfrac{1}{L_1} + \dfrac{1}{L_2} + \dfrac{1}{L_4} ... + \dfrac{1}{L_m}\right) \\ \qquad\qquad\qquad ... \\ \dfrac{dL_m}{dt} = k_{L_m} L_m (N - L_m) \left(\dfrac{1}{L_1} + \dfrac{1}{L_2} + \dfrac{1}{L_3} ... + \dfrac{1}{L_{m-1}}\right) \end{cases}$$

This describes the relationship between m interacting languages, which could be numerically simulated using Euler's Method in order to estimate future patterns. More details with this calculation would be discussed later on in 4.2.3.

D. Influence/Learning Matrix and Vector Representation

One may notice that in the previous models, there is no quantitative consideration of the exact influencing factors that may contribute to the motive for people to pick up a second language and become a Non-native Speaker, such as immigrations, education requirements, business intercontinental transactions and tourisms. These factors were all generalized into coefficients without further breakdown. Therefore, this final model on the Total Speaker patterns would be analyzing each of the main factors laid out in the background introduction, as well as some newly added factors from originality. As always, we would start from the simplest case of two languages (and two Language Zones).

Since the influencing factors are broken down, their individual effect might act on or result from the Native Speakers and the Non-native Speakers differently; this is why the following vector is designed to indicate the combination of those amounts as a "speaker profile" for each of the 2 languages:

$$[S_{1A} \quad S_{1B} \quad S_{2A} \quad S_{2B} \quad 1]^T$$

S_{1A} is the amount of Native Speakers for language A, and S_{2A} is the amount of Non-native speakers for language A. S_{1B} and S_{2B} are similarly defined for language B. 1 allows independent constant terms to be added. By considering this status as a vector, changes brought by each influence factor over a time unit could be represented by 5×5 matrices left multiplied to this vector successively. This allows as to break down each of the factors as a matrix and combining them simply by multiplying them together to form one complex 5×5 matrix that takes all factors into account. We would now investigate the Influence/Learning Matrices separately for each influencing factor. Since we are looking at the motives to learn, it is necessary to keep our point of view from the side of the potential learner to analyze the influence.

*One should realize that the coefficients such as k_A varies with each sub-model (matrix).

Immigrants and Assimilation

Immigrants, when moving to another Language Zone, are forced to learn the local language in order to fulfill their basic living needs. However, the speed at which they "cave" to learn is dependent on the population of the Native Speakers in this zone. Consider situations when 100 immigrants moving into a country with 150 people and when the same immigrants moving into a country with 1000000 citizens. The 100 immigrants would be more separated in the latter situation, and in comparison more surrounded by native speakers, and hence more likely to learn to become a Non-native Speaker. Therefore, the following matrix is proposed to describe this change:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ k_A I_{B\to A} & 0 & 1 & 0 & 0 \\ 0 & k_B I_{A\to B} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} S_{1A} \\ S_{1B} \\ S_{2A} \\ S_{2B} \\ 1 \end{bmatrix}$$

In the above matrix, $I_{A\to B}$ denotes the number of immigrants moving from A to B, and $I_{B\to A}$ is similarly defined. Notice how the term $k_A I_{B\to A}$ is placed in the first column to be multiplied with S_{1A} to increase S_{2A}. This is an analogy to the term $aS(n)I(n)$ in the SIR model in 4.2.2, in which the number of potential patients is proportionally correlated to the infected population according to the **Size-Influence Principle**.

Business Relation

In order to open up new foreign markets or cooperate deeper with current markets, many businesspeople choose to learn the native language of their foreign customers, thus becoming a Non-native Speakers of that certain language. The language they choose to learn depends on the population of Language Zones, which is related to the market size and hence potential number of transactions. They therefore have more incentive to learn the native languages of the bigger markets according to the **Size-Influencing Principle**. On the other hand, in order to solidify the cooperation with the existing customer base, a businessperson will also be more likely to choose to learn the language which he or she currently needs to use the most.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ k_A & 0 & 1 & 0 & k_A{}'BR_{AB} \\ 0 & k_B & 0 & 1 & k_B{}'BR_{AB} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} S_{1A} \\ S_{1B} \\ S_{2A} \\ S_{2B} \\ 1 \end{bmatrix}$$

In the above matrix, BR_{AB} denotes the number of business relations between Language Zone A and Language Zone B. Deepening current cooperation is independent on the market size hence put in the fifth column for constants, and k_A, k_B are related to the native market size by the **Size-Influence Principle**, hence put in the first and second column.

Tourism

When people largely visit another place, the local people, if of a relatively small size and especially those who depend highly on tourism, would consider to learn a second language to have better business opportunities. In such occasions, completely opposite from the immigrant, it is the visited place that would learn the dominant languages of the visitors.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 + k_A k_A' & 0 & k_A T_{A \to B} \\ 0 & 0 & 0 & 1 + k_B k_B' & k_B T_{B \to A} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} S_{1A} \\ S_{1B} \\ S_{2A} \\ S_{2B} \\ 1 \end{bmatrix}$$

As shown in the last column above matrix, the effect of the number of tourists visiting is independent from the rest of the population size, and hence multiplied by the constant 1. The terms $k_A k_A'$ and $k_B k_B'$ are for those whose native language isn't A but has a second language A such that learning A could also help the local tourism workers to communicate with them.

Translation Technology

The ever-developing translation technology decreases people's desire to learn a new language. As depicted in the famous sci-fi *The Hitchhiker's Guide to the Galaxy*, there is a species of fish called *Babel Fish* that lives in people's ears, and is able to transform foreign languages into people's native languages automatically and perfectly. People do not need to learn a new language to communicate at all because of the existence of *Babel Fish*. It can be inferred that translation technology will restrain the number of new Non-native Speakers, but will not decrease the existing number of Non-native Speaker, since technology cannot let them forget what they already know.

We would define the technology factor as Te between 0 and 1, where 0 means none translation technology and 1 means *Babel Fish*'s perfect translation.

Effect on "Tourism" Matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 + (1 - Te) k_A k_A' & 0 & (1 - Te) k_A T_{A \to B} \\ 0 & 0 & 0 & 1 + (1 - Te) k_B k_B' & (1 - Te) k_B T_{B \to A} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Notice how translation technology only has effect on the new language learners (changes) but not the original portion of Non-native Speakers.

Used/Taught in Education/School

When a foreign language is used or taught in schools, students will have the motive to learn it to get good grades. The increased number of Non-native Speakers of a language is solely related to the number of students learning it successfully, but not to any existing number of speakers, hence put in the last column as constant.

$$\begin{bmatrix}1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & E_A\\ 0 & 0 & 0 & 1 & E_B\\ 0 & 0 & 0 & 0 & 1\end{bmatrix} \times \begin{bmatrix}S_{1A}\\ S_{1B}\\ S_{2A}\\ S_{2B}\\ 1\end{bmatrix}$$

Demographic Changes

For Native Speakers, their offspring will immediately count as a Native Speakers; however, the offspring of the non-native speakers, their native language would be dependent on the language in the place that they are born, hence not inherited.

$$\begin{bmatrix}1+b-d & 0 & 0 & 0 & 0\\ 0 & 1+b-d & 0 & 0 & 0\\ 0 & 0 & 1-d & 0 & 0\\ 0 & 0 & 0 & 1-d & 0\\ 0 & 0 & 0 & 0 & 1\end{bmatrix} \times \begin{bmatrix}S_{1A}\\ S_{1B}\\ S_{2A}\\ S_{2B}\\ 1\end{bmatrix}$$

Language Used in Government Organization

When a foreign language is used in government organizations or set as the official language by government like English in India, the purpose for the citizens to learn that language is to find a job issued by the government or deal with related issues. People's desire to learn the language is not related to the current status of the language and its Language Zone, and hence put in the the constant column.

$$\begin{bmatrix}1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & k_A P_{BA}\\ 0 & 0 & 0 & 1 & k_B P_{AB}\\ 0 & 0 & 0 & 0 & 1\end{bmatrix} \times \begin{bmatrix}S_{1A}\\ S_{1B}\\ S_{2A}\\ S_{2B}\\ 1\end{bmatrix}$$

In the matrix above, P_{AB} is a dummy variable that is equal to one when Government A uses Language B, and is equal to zero otherwise.

Pop Culture Output

Pop culture will also influence the number of Non-native Speakers of a language. The number of followers that a certain pop culture has does not depend on the number of Speakers of the language used in that culture. Consider the example of India and Korea, Korean culture has a much larger fan base than Indian culture such as K-pop and entertainment shows such as *Running Man*, even though Korean has a much smaller Language Zone and size of speakers than Hindi. Therefore, the cultural output is independent with the size of native or non-native speakers.

$$\begin{bmatrix}1 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & k_A C_A\\ 0 & 0 & 0 & 1 & k_B C_B\\ 0 & 0 & 0 & 0 & 1\end{bmatrix} \times \begin{bmatrix}S_{1A}\\ S_{1B}\\ S_{2A}\\ S_{2B}\\ 1\end{bmatrix}$$

Therefore, after analyzing the above 8 factors in influencing the combination of native and non-native speakers, the product of the matrices in order of greatest influence to the smallest, could represent the complex and comprehensive change in the speaker population combination represented by the vector. To produce a long term prediction, the further calculations of powering the matrix is required; considering the product matrix undergone diagonalization by eigenvalues and eigenvectors to be added a power to, the problem of massive calculation scale could soon be resolved.

For cases with more languages, discussing the interaction between either two of them one-by-one and leaving the rest of the matrix identical to identity matrix would extend this method to higher degrees.

4.2.4. Calculations and Results

The models we have presented on modeling the distributions of various language speakers over time could be seen as our approach to Problem A in Part I; when more coefficients and data are available, the Influence Matrix approach is recommended; otherwise, the SIR model could generate a relatively accurate model with Euler's Method in the short term.

As for Problem B in Part I, due to the lacking of related constants, some of the models might be inconsistent and unstable; this paper has combined the results of two models to tackle this obstacle.

We have mentioned in the end of the Multiple Language General Model, the constant k is the one constant that needs to be determined which is related to the factor of growth; however, given past speaker population data, we could simply use the logistic model to generate the power coefficient $Nk_1 + b - d$ to be the factor of growth, and then plug into the Multiple Language General Model to do predictions. Using features of Excel and past data from *Ethnogue*, we have predicted the future pattern for Total Speakers in 50 years to be: (in millions)

Language	2017 Total	2017 Ranking	Growth Factor	2067 Total	2067 Ranking
Chinese	1284	1	0.00670	1346.227	2
Spanish	437	2	0.13618	1110.293	3
English	372	3	0.2138	1572.558	1
Hindi	260	5	0.04814	362.170	5
Arabic	295	4	0.0625	453.232	4
Portuguese	219	7	0.04274	293.430	7
Bengali	242	6	0.03392	305.271	6
Russian	154	8	0.02474	182.201	9
Japanese	128	9	0.01182	138.201	10
Javanese	84.4	10	0.011816	109.077	11
German	76.8	12	0.03924	92.806	14
Korean	77.2	11	0.02922	94.161	13
French	76.1	13	0.03064	184.536	8
Telugu	74.2	14	0.13622	94.437	12

As we could clearly see, French is the language breaking into the new Top 10 at number 8, and Javanese is being pushed out of the top 10; another noticeable phenomenon is that English has become the language with the most speakers in the world exceeding Spanish and Chinese already. The native speaker trend, in contrast was quite peaceful without replacement.

4.3. Geographical Distribution Model

4.3.1. Global Population and Human Migration Patterns

The question requires us to first predict the “global population and human migration patterns” for the next 50 years. Since these are the minor supporting models, this paper has adopted the logistic model of population and the gravity model of migration to produce these necessary numbers. The famous logistic model is quite self-explanatory and well applied, so the main focus here would be on the Gravity Model of Migration.

“The key to the gravity model is the relationship between migration and distance,” as claimed in a paper titled *Modeling Migration* produced by Greenwood from Colorado University. Seeing as how the gravity of “pulling” immigrants population over is inversely proportional to the square of distance, the following model is hence developed:

$$I_{B\to A} = k\frac{GDP_{per\ capita\ A} - GDP_{per\ capita\ B}}{GDP_{per\ capita\ B}(d^2)} \times n_B$$

In this case, $I_{B\to A}$ is the number of immigrants moving from B to A. Using the assumption that people would tend to move to places to seek higher income and better lives, the percentage difference in GDP per capita is also used as a factor to determine the portion of the total population n_B that decides to migrate.

4.3.2. Factors of Geographical Distribution and Algorithm

Having done the above preparation work, now this paper would officially start to investigate the geographical distribution of language speakers. In order to simplify the problem, only the amount of Total Speakers is investigated; the more accurate decomposition could be easily done in a similar manner or use the matrices when data are available.

Using Python to realize Euler’s Method, we have made the following steps with each of the cities (example here: city A_1 speaking language A)over every time unit:

1. Demographic Change by Logistic Regression

   Both the population and Native Speaker Populations are changed, hence:

$$P_{A_1}{}' = P_{A_1} + k, S_{A_1A}{}' = S_{A_1A} + k$$

2. Learning Second Languages

   The Speaker Populations would be changed, hence by the General Multiple Language Model, we have for any language X:

$$S_{A_1X}{}' = S_{A_1X} + \frac{k_X L_X(N - L_X)}{L_A}$$

3. Immigrations

   Using knowledge from the Immigration Matrix and the Gravity Model,

$$I_{X_i\to A_i} = k\frac{Income_{A_1} - Income_{X_i}}{Income_{X_i} d^2} \bullet P$$

   Hence the speaker population changes locally moving in and out,

$$S_{A_1A}{}' = S_{A_1A} + \sum I_{X_i\to A_1} - \sum I_{A_1\to X_i}$$

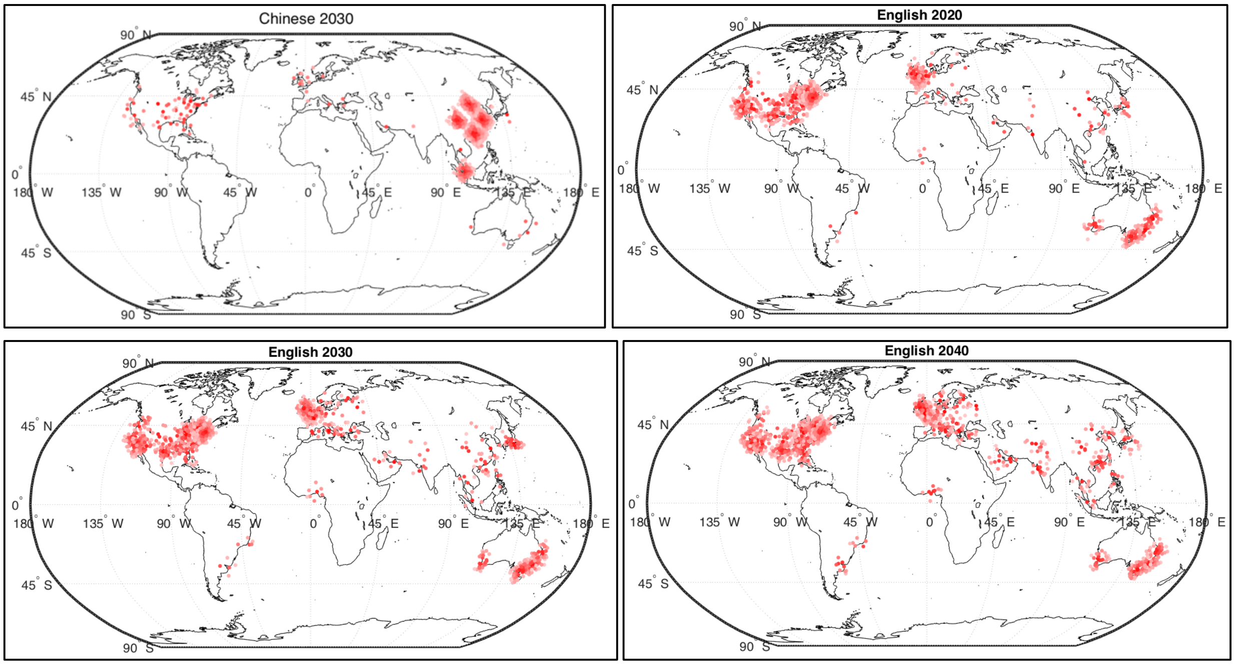
$$S'_{A_1X} = S_{A_1X} + \sum(I_{X_i\to A_1} - I_{A_1\to X_i}) + \sum_{Y=B,C,..}\sum_{i=1}^{n}(I_{Y_i\to A_1}\frac{S_{Y_iX}}{P_{Y_i}} - I_{A_1\to Y_i}\frac{S_{A_1Y}}{P_{A_1}})$$

4. Recalculate Total Speaker Number to apply Size-Influence Principle,

$$L'_A = \sum_X \sum_i S_{X_i A}\,, P' = P + \sum I_{X_i \to A_1} - \sum I_{A_1 \to X_i}.$$

4.3.3. Calculations and Results

Having predicted the total speaker for the major places and put in a giant matrix, we have chosen to use graphical display to more directly present the result of our prediction.



The above four diagrams respectively demonstrate the language Chinese's geographical distribution in year 2030, and English's geographic distribution change over time from 2020 to 2040. The red dots indicate the number of speakers in that region, and the deeper the color is, the denser the speakers are. Please make sure to zoom in to view the details of the maps. As we can from the three English distribution diagrams, as time passes, English has been spread to not only the Language Zone of UK, US, Canada and Australia, but also other parts of the world like China, South America, Japan, Europe, India, and North Africa. Over time, the density of the dots in these new parts seem to increase, indicating that there are more and more learners of English across the globe. This is essentially our presentation of our prediction for geographic distribution for Part I C.

4.4. Site Selection Model

4.4.1. Evaluation Components of City Combination

In order to evaluate any combination of cities to be selected for offices to be built in, an evaluation index needs to be constructed first as the target of programming. After considering various factors to be taken into account into the site selection process, we have eventually landed upon the following three components, Economic Component, Language Component and Location Component, divided into five elements to add up to make a final decision.

1. **Economic Component: Market Size**

The market size of the city that the office is in is very important because it primarily determines the business's primary size of opportunity. The company's client base, raw material purchase, and labor hiring are all basing upon the market size to determine their availability and price. Hence this is very essential to the discussion; this element would be quantitatively measured and maximized in terms of the GDP of the city.

2. **Economic Component: Labor Cost**

The cost of building a new office, except for the common bases of housing and office material purchases, bases on majorly the cost of the labor. Especially in this multinational company in the service sector, expensive labor could be a heavy burden to its operating cost; this element would be measure in terms of average wage/income of the city.

3. **Language Component: Speaker Composition**

Combining with the language research having been done, the following process is taken in determining the number and kinds of languages used in the office with variable criteria 95%:

Taking the Locally Prominent (Highest Spoken Percentage) Language, until Total Speaker Coverage (total percentage of people able to be communicated with using either of the languages used) reaches 95%.

After selecting the languages used for every office, the following criterion is used for the language component:

$$\frac{\%\ of\ English\ Speakers}{(number\ of\ languages-1)^2}$$

% of English Speakers is for the benefit of hiring since English as a language is fixed to be used in all companies; the less extra number of languages, the less complicated work is going to be inside the office without lingual disorder, hence put in the denominator.

4. **Location Component: Economic Centeredness**

This criterion is divided into two steps. First of all, this paper uses the K-means algorithm to divide the main cities into portions by considering them as points on the earth surface with weights proportional to their GDP. The main goal is to select one city from each region such that all main cities around the globe are covered by the company. This could also greatly reduce the scale of calculation.

When the above criterion is reached, one needs to calculate the following term for each of the offices in the evaluated combination:

$$\sum\sum\frac{GDP(for\ each\ city)}{d^2+1}$$

This is to consider how well the companies are reaching out to every main city, and hence the total available market shares and company size its location centeredness could allow. The further away it is from the cities, the less share it would get of that city's market measured by GDP.

5. **Location Component: Workload Bearing**

This component is derived from the thought that each office should be in charge of a similarly sized market, such that it wouldn't be too big a burden for any of them. Taking the reciprocal of the standard deviation of the total market size for each office divided using the Voronoi Diagram, we could get our measuring criteria for this element.

4.4.2. Analytic Hierarchy Process (AHP)

The problem now becomes how to combine the above criteria. Since none of the two criteria are interdependent, we have chosen to use a linear combination of the normalized values of each criteria. The coefficients before each term is hence very important. When getting in touch with the real client company, the relative importance of these business factors could be determined based on their economic goal or preference. However, currently we need to base on empirical values and our interpretation of the values. Even though the objectivity is not guaranteed, in reality, the actual company would also make subjective judgments about the relative importance between factors, which justifies our choice of the AHP model. Please realize that these constants could be altered to generate different results easily, but the same model would still be applied.

Hence, the Analytic Hierarchy Process is introduced. Using the following series of importance, the following comparison matrix is developed:

(Important) GDP, Centeredness, Language, Labor Cost, Workload Bearing (Least Important)

	Language	Centeredness	Labor Cost	Market Size	Workload
Language	1	0.33	1.33	0.25	3
Centeredness	3	1	5	0.5	8
Labor Cost	0.8	0.17	1	0.25	1.5
Market Size	4	2	5	1	10
Workload	0.5	0.14	0.5	0.09	1

This comes with a CI value of 0.033, and a CR value of $0.029 < 0.1$, therefore the following coefficients are valid:

Language 0.11, Centeredness 0.31, Labor Cost 0.08, Market Size 0.46, Workload Bearing 0.04.

4.4.3. Calculations and Results

Having selected the biggest 100 cities in the world economically according to their GDP, we have been evaluating how the combination of company would become after adding any one of them. The following are the results that we obtained.

Short Run Selection (around 10-year span):

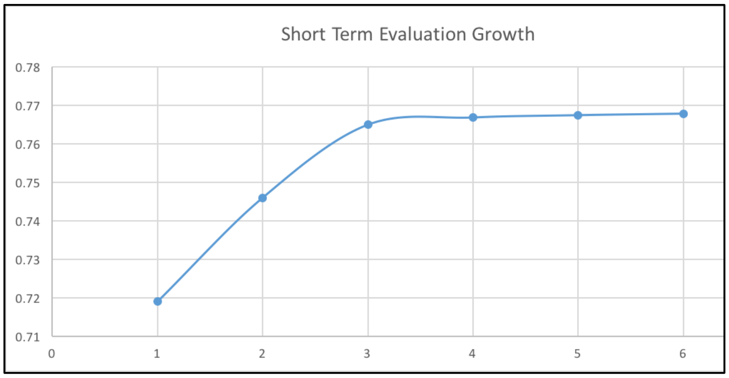
Office Location City	Language Used	Evaluation Coefficient
Amsterdam, Netherland	English, Spanish	0.719
London, Britain	English, Arabic	0.746
Paris, France	English, French	0.765
Osaka, Japan	English, Japanese	0.7669
Beijing, China	English, Chinese	0.7675
Sydney, Australia	English, Arabic	0.7679

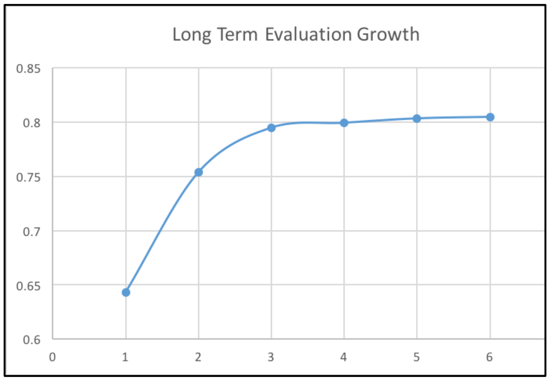
Long Run Selection (around 30-year span):

Office Location City	Language Used	Evaluation Coefficient
Paris, France	English, French	0.6433
Osaka, Japan	English, Japanese	0.7538
Beijing, China	English, Chinese	0.7947
Toronto, Canada	English, French	0.7993
Delhi, India	English, Hindi	0.8033
Guangzhou, China	English, Chinese	0.8047

The difference between the two cases actually makes quite a lot of sense. As we could see, in the long term, there are more Asian countries selected whose native language is not English. They may have a disadvantage because of that in the short run, but in the long run they have a lot of time to catch up on their language and, because of their massive population, be ahead economically. Therefore, it is beneficial to select offices in those areas in the long term.

The order (or priority) of selection is from high to low; one may wonder why the top cities are the first ones to be locked in given they have lower values of the evaluation coefficient. Well, the reason is that these new offices gives the best improvement on the coefficient. Obviously, the more office there are, the higher the coefficient would be because of the higher GDP, the higher centeredness, etc. However, it would tend to be more difficult to raise upon a very high level because the centeredness and the language components couldn't be raised by much, while the rest of the cities to choose from are hardly perfect (if they are, they would be the priority already). The following two diagrams are drawn to show this diminishing return effect:





As we could clearly see, the evaluation coefficient has been increasing ever slower for every extra office. This means that when the company have an increase expectation for this evaluation on how well the companies are working together, this prediction could be a potential stopping sign for the company: even choosing the best possible city in the world wouldn't improve the evaluation by much. This is the criteria for which occasions of building less than 6 companies be planned.

For example, if the expected time span is 10 years, and the company expects an at least 0.001 growth for every new office, it probably should have decided to stop at the first 4 offices.

5. Sensitivity Analysis

First of all, we would analyze the stabilized result from the basic discrete SIR model.

$$I(n) = \frac{Nbn_i}{Nd - k}$$

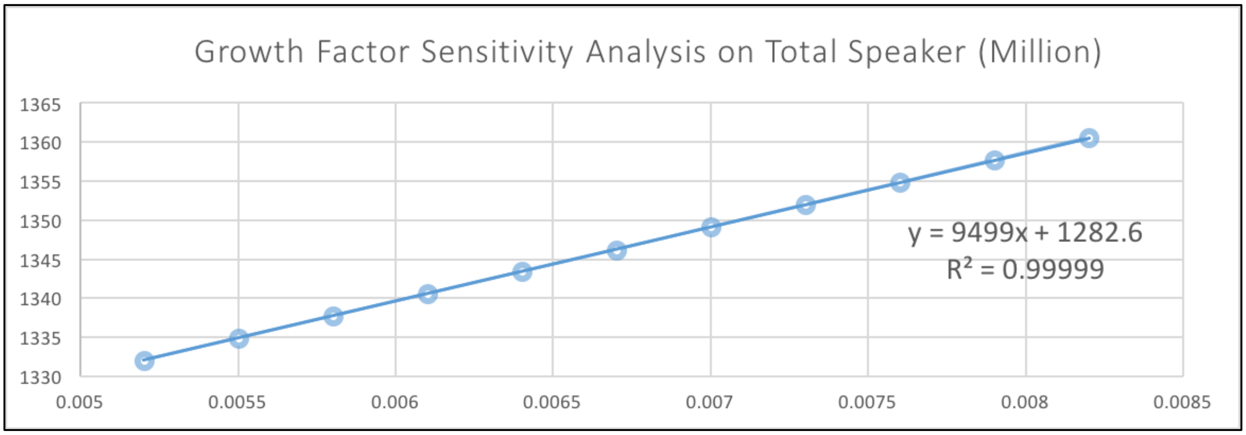
Analysis for the change in b and n_i is quite simple; the linear relationship guarantees a relative elasticity of 1, proven by the following example:

$$\frac{dI}{db} \times \frac{b}{I} = \frac{Nn_i}{Nd - k} \times \frac{b(Nd - k)}{Nbn_i} = 1$$

Analysis for d (death rate):

$$\begin{aligned}\frac{dI}{dd} \times \frac{d}{I} &= -\frac{Nbn}{(Nd - k)^2} \times N \times \frac{d}{I} \\ &= -\frac{Nd}{Nd - k} \\ &= -1 - \frac{k}{Nd - k} \approx -1.8902 < -1\end{aligned}$$

Which means there is a slightly stronger and more elastic negative relation than b. Then, the growth factor resulted from logistics regression is analyzed against, but since the answer if derived from Euler's Method, there's no analytical but only numerical data, hence only using the following diagram to analyze:



The high R^2 value indicates a strong linear relationship is suggested in the domain, which makes the model much more predictable and applicable in a wider range in this consistent sensitivity.

6. Strengths, Weaknesses and Improvements

Strengths:

1. Multiple models of predicting total speaker trend are proposed, with each one basing off others' results, gradually breakdown and approach this complicated issue. Each also provides a different option for different availabilities and accuracies of data source.
2. The innovative comparison and connection between languages and diseases makes SIR model applicable in this special context, and helped to conclude the size-influence principle which is then applied many times as an essential idea.
3. The graphical and theoretical analysis of the stable points in the altered SIR model is very convincing and generates results quickly.
4. Using matrices to separating the effect of each of the influencing factors is a quick and direct way of breaking down the problem. The immigrant matrix also contributed to later discussions in geographical distributions. It would also be much easier to add new factors by creating a new matrix. The vector representation of the composition of speaker is not only mathematically elegant but also has great real-life meanings.
5. The graphic representation of results is direct and easy to understand by non-experts.
6. The evaluating model of city combinations considers multiple aspects including language, economy, and locations to generate a comprehensive result.
7. The location component combines various geographical and geometrical elements such as the K-means algorithm and Voronoi Diagram to make the model make geographical sense and also elegant mathematically.
8. AHP method and the allowance for long and short run along with the consistent sensitivity analysis makes the model adaptable in various situations.

Weaknesses:

1. The small data set has forced us to go on Ethnologue to find out some old data, while linguistic data are usually highly unreliable.
2. The multiplication of matrices requires the computer to operate multiplications of around $\binom{23}{2}\times 50\times 8=5\times 10^6$ matrices each of size 49×49, which is too much calculation with high risks of inaccuracies and mistakes.

Improvements:

1. Use the multiplication of matrices into writing a complex differential equation to instead make predictions might be the solution, yet much less elegant.
2. Decompose the coefficients that represent the multi-factor situation to predict the future trend with a more detailed existing model such as the SIR model.

7. Conclusion (Memo to Chief Operating Officer)

Dear Chief Operating Officer:

Greetings! Having studied and considered your needs as a large multinational service company, we have made the following suggestions to you to open new offices:

If choose to open 6 new offices:

Short Run Selection (around 10-year span):

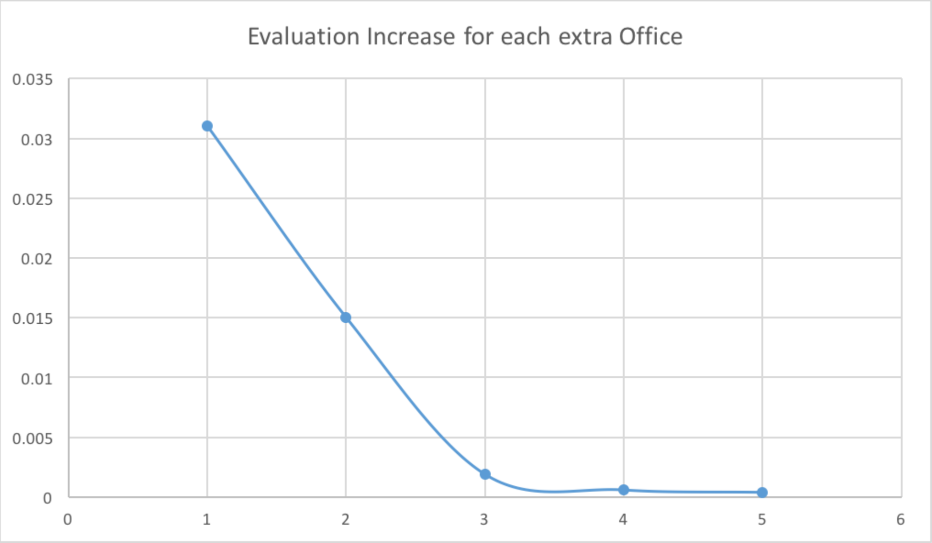
Office Location City	Language Used
Amsterdam, Netherland	English, Spanish
London, Britain	English, Arabic
Paris, France	English, French
Osaka, Japan	English, Japanese
Beijing, China	English, Chinese
Sydney, Australia	English, Arabic

Long Run Selection (around 30-year span):

Office Location City	Language Used
Paris, France	English, French
Osaka, Japan	English, Japanese
Beijing, China	English, Chinese
Toronto, Canada	English, French
Delhi, India	English, Hindi
Guangzhou, China	English, Chinese

The order (or priority) of selection is from high to low; that is to say, to build office in Amsterdam first in the short term, and in the long run build office in Paris first. You could also change the coefficients of the constants in the AHP model, whose subjectivity is suitable to make oriented goals for companies like this.

The diminishing marginal return for each extra office built means that if your company would consider its expectation for the growth in the evaluation coefficient for each extra company, decisions when less than 6 extra offices to be built could also be determined by using the following diagram:



In the above diagram, the diminishing return effect is drawn for each extra office. According to the expectation of the company, you could freely choose where to stop investing in building new offices when the increase in evaluation is smaller than your expectation. In which case, the prioritized (top of the tables) should be built first.

8. Bibliography

Cost of Living, 2018, www.numbeo.com/cost-of-living/.

Ethnologue, www.ethnologue.com/statistics/size.

"2018 World Population by Country." *2018 World Population by Country*, worldpopulationreview.com/.

"Contact CIA." *Central Intelligence Agency*, Central Intelligence Agency, www.cia.gov/Library/publications/the-world-factbook/fields/2098.html.

Giordano, Frank R., et al. *A first course in mathematical modeling*. Brooks/Cole Cengage Learning, 2014.

Greenwood, Michael J. "Modeling Migration." *Modeling Migration*, Colorado University, 2005, www.colorado.edu/Economics/courses/spring12-4292-001/Modeling%20Migration.pdf.

9. Appendix

a. Diagram Generating Program Script

LoadData.m

```
%% This script loads essential data of cities from an excel file.

clear;
clc;

[num,txt,~] = xlsread('../Data/CityEconStats.xlsx','data');

P18 = num(2:end,1);
P15 = num(2:end,2);
P10 = num(2:end,3);
latd = num(2:end,5);
lond = num(2:end,4);
lon = deg2rad(lond);
lat = deg2rad(latd);
income = num(2:end,6);

en2020 = num(2:end,10);
en2030 = num(2:end,11);
en2040 = num(2:end,12);
zh2030 = num(2:end,13);

city = txt(2:end,1);

clear num;
clear txt;

% Prepare for regression
X = [10;15;18];
pop = [P10,P15,P18];
```

RegressPop.m

```
%% Performs linear regression on city population.
reg_result = zeros(90,3);
X = [ones(3,1),X];
for i = 1:size(pop,1)
    [b,~,~,~,stats] = regress(pop(i,:)',X);
    reg_result(i,1:2) = b';
    reg_result(i,3) = stats(1);
    clear b;
    clear stats;
end
% x * reg_result(1,2) + reg_result(1,1)
```

DrawMap.m

```
%% This script performs the task of data visualization given the result
calculated by the model.

%% English 2020
l=length(city);
h=worldmap('World');
geoshow('landareas.shp','FaceColor', 'white');
```

```
for k=1:l
    for j=1:en2020(k)/5e4
        offset1 = normrnd(0,5);
        offset2 = normrnd(0,5);
        value = (abs(offset1)+abs(offset2))/10;
        if value <= 0.8
            plotm(lond(k)+offset1,latd(k)+offset2,'Marker',
'.','MarkerSize',15,...
                'Color', [1 value value]);
        end
    end
    % plotm(lond(k),latd(k),'Marker', '*','MarkerSize',10,'Color', [1 0
0]);
    display(k,' finished')
end
title('English 2020');
saveas(gcf,'en20.png')
% textm(lond,latd+5,city);

%% English 2030
h=worldmap('World');
geoshow('landareas.shp','FaceColor', 'white');
for k=1:l
    for j=1:en2030(k)/5e4
        offset1 = normrnd(0,5);
        offset2 = normrnd(0,5);
        value = (abs(offset1)+abs(offset2))/10;
        if value <= 0.8
            plotm(lond(k)+offset1,latd(k)+offset2,'Marker',
'.','MarkerSize',15,...
                'Color', [1 value value]);
        end
    end
    % plotm(lond(k),latd(k),'Marker', '*','MarkerSize',10,'Color', [1 0
0]);
    display(k,' finished')
end
title('English 2030');
saveas(gcf,'en30.png')
% textm(lond,latd+5,city);

%% English 2040
h=worldmap('World');
geoshow('landareas.shp','FaceColor', 'white');
for k=1:l
    for j=1:en2040(k)/5e4
        offset1 = normrnd(0,5);
        offset2 = normrnd(0,5);
        value = (abs(offset1)+abs(offset2))/10;
        if value <= 0.8
            plotm(lond(k)+offset1,latd(k)+offset2,'Marker',
'.','MarkerSize',15,...
                'Color', [1 value value]);
        end
    end
    % plotm(lond(k),latd(k),'Marker', '*','MarkerSize',10,'Color', [1 0
0]);
    display(k,' finished')
end
title('English 2040');
saveas(gcf,'en40.png')
% textm(lond,latd+5,city);
```

```
%% Chinese 2030
h=worldmap('World');
geoshow('landareas.shp','FaceColor', 'white');
for k=1:l
    for j=1:zh2030(k)/5e4
        offset1 = normrnd(0,5);
        offset2 = normrnd(0,5);
        value = (abs(offset1)+abs(offset2))/10;
        if value <= 0.8
            plotm(lond(k)+offset1,latd(k)+offset2,'Marker',
'.','MarkerSize',15,...
                'Color', [1 value value]);
        end
    end
    % plotm(lond(k),latd(k),'Marker', '*','MarkerSize',10,'Color', [1 0
0]);
    display(k,' finished')
end
title('Chinese 2030');
saveas(gcf,'zh30.png')
% textm(lond,latd+5,city);
```

b. Geographic Distribution Calculation Script

```
import xlrd

#Open Distance Sheet
dist = xlrd.open_workbook('Distance(1).xlsx')
distTable = dist.sheets()[0]
cityName = distTable.col_values(0) #Create a list with all the names of cities



#Get access to GDP data
raw = xlrd.open_workbook('CityEconStats1.xlsx')
rawTable = raw.sheets()[5]
GDP20 = rawTable.col_values(22)
GDP30 = rawTable.col_values(25)
GDP40 = rawTable.col_values(28)
GDP12 = rawTable.col_values(48)
GDP14 = rawTable.col_values(49)
GDP16 = rawTable.col_values(50)
GDP18 = rawTable.col_values(51)
GDP22 = rawTable.col_values(52)
GDP24 = rawTable.col_values(53)
GDP26 = rawTable.col_values(54)
GDP28 = rawTable.col_values(55)
GDP32 = rawTable.col_values(56)
GDP34 = rawTable.col_values(57)
GDP36 = rawTable.col_values(58)
GDP38 = rawTable.col_values(59)

#Get the distance of Shanghai and New York to other cities
SHD = distTable.row_values(63)
NYD = distTable.row_values(77)
```

```
#The central index for one city
def central(city):
    cityList = [] #determine the city's reposible customers (other cities)
    sum = 0
    CityD = distTable.row_values(cityName.index(city))
    for i in range(1,len(SHD)):   #For different city, which one of shanghai,'city',
and new york is better
        if (1/(SHD[i]+1)) < (1/(CityD[i]+1)) and (1/(NYD[i]+1)) < (1/(CityD[i]+1))
and i != cityName.index(city):

cityList.append((GDP12[i]+GDP14[i]+GDP16[i]+GDP18[i]+GDP20[i])/(1+CityD[i]))
            sum += (GDP12[i]+GDP14[i]+GDP16[i]+GDP18[i]+GDP20[i])/(1+CityD[i])
        else:
            if (1/(SHD[i]+1)) < (1/(NYD[i]+1)) and cityName[i].find('Shanghai') == -
1:
                sum += (GDP12[i]+GDP14[i]+GDP16[i]+GDP18[i]+GDP20[i])/(1+SHD[i])
            elif (1/(NYD[i]+1)) < (1/(SHD[i]+1)) and cityName[i].find('NewYork') ==
-1:
                sum += (GDP12[i] + GDP14[i] + GDP16[i] + GDP18[i] + GDP20[i]) / (1 +
NYD[i])
            else:
                continue
    return sum, cityList

#the mean for the central index
total = 0
cityList = []
for i in range(1,len(cityName)):
    total += central(cityName[i])[0]
    cityList.append(central(cityName[i])[1])
mean = total/89
print(mean)

#find the variance of the central index
def variance(city):
    sumVar = 0
    for i in range(0, len(cityName) - 1):
        var = (cityList[cityName.index(city)-1][i] - mean)**2
        sumVar += var
    return sumVar

#print out the variances
for i in range(1,len(cityName)):
    print(variance(cityName[i]))
```

c. Evaluating Criteria Calculation Script

```
import xlrd,numpy
import math as m
#Derive the excel sheets need
raw = xlrd.open_workbook('CityEconStats.xlsx')
coeff = xlrd.open_workbook('linear regression.xlsx')
coTable = coeff.sheets()[0]
rawTable = raw.sheets()[2]

#A List with all existing languages in the database
TotalLangList = ['English','German','French',"Chinese","Hindi","Arabic","Spanish"
```

```
    ,"Russian","Lahnda","Bengali","Japanese","Korean","Portuguese"]

#The list with  all the growth rate of all country
kList0 = coTable.col_values(2)
del(kList0[0])
kList = []
#float 转int
for k0 in kList0:
    k = int(k0)
    kList.append(k)
print(kList)


#a dictionary for all the coefficents for langauges
cDict = {'Chinese':0.00003348,
'English':0.00004276,'Spanish':0.0006809,'Hindi':0.000247,

'Arabic':0.0003125,'Portuguese':0.0002137,'Bengali':0.0001696,'Russian':0.0001237,

'Japanese':0.00005908,'Lahnda':0.0005434,'German':0.0001461,'Korean':0.0001532,
         'French':0.0006811}




#City Names
cityName = rawTable.col_values(0)
del(cityName[0])
del(cityName[90])
print(cityName)

#Native Language Type for each country
langType = rawTable.col_values(1)
del(langType[0])
print(langType)

#Each country's population initially
OGpop = rawTable.col_values(4)
del(OGpop[0])
OGpop[68] = "".join(OGpop[68].split())
OGpop[68] = float(OGpop[68].replace(',',''))
print(OGpop)

#Each country's longtitude and latitude
x = rawTable.col_values(5)   #longitude
z = rawTable.col_values(6)   #latitude
del(x[0])
del(z[0])


#Each country's citizens' average income, derived from excel sheet
income = rawTable.col_values(7)
del(income[0])

#The distance between two cities
def d(x1,x2):
    d = 6371*m.acos(m.cos(x1[4]*m.pi/180)*m.cos(x2[4]*m.pi/180)*
                    m.cos((x1[3]-
x2[3])*m.pi/180)+m.sin(x1[4]*m.pi/180)*m.sin(x2[4]*m.pi/180))
    return d

#A list that consists of  [city name, language type, longitude, latitude]
rawList = []
for i in range(90):
    rawList.append([cityName[i],langType[i],OGpop[i],x[i],z[i],income[i]])
print(rawList)

#A formula for immigration from city x to city y
def I(x,y):
    Ixy = 0.01*(y[5]-x[5])*y[2]/(x[5]*d(x,y)**2)
    return Ixy

#Building a matrix that consist all the information for a city
matrix = []
subMatrix = []
CoPop = 0 #To determine whether a language is a speaker's native language
for lang in TotalLangList:
    for city in cityName:
        i = cityName.index(city)
        if (rawList[i][1].find(lang)) == 0:
            CoPop = 1
        else:
            CoPop = 0

subMatrix.append([city,lang,CoPop*rawList[i][2],rawList[i][3],rawList[i][4],rawList[
i][5]])
    matrix.append(subMatrix)
    subMatrix = []
print(matrix)


LangT = [] #Total speakers for different language
LangN = [] #To match the langauge
TotalPop = 0 #Total population
for sub in matrix:
    sum = 0
    lang = sub[0][1]
    for subsub in sub:
        sum += subsub[2]
    LangT.append(sum)
    TotalPop += sum
    LangN.append(lang)
print(TotalPop)
print(LangT)
print(LangN)

#Total population for each country
def P(x):
    sum = 0
    for i in range(len(matrix)):
        sum += matrix[i][x][2]
    return sum
```

```
for year in range(10):
#No.native speakers increase/decrease through time
    for sub in matrix:
        j = matrix.index(sub)
        Co = 0 #determine whether a language is the speaker's native language
        for subsub in sub:
            i = sub.index(subsub)
            if (rawList[i][1].find(subsub[1])) == 0:
                Co = 1
            else:
                Co = 0
            subsub[2] = int(subsub[2]) + Co*kList[i] #the effect of having babies on 
number of speakers in a city
            LangT[matrix.index(sub)] += Co*kList[i] #the effect of having babies on 
number of speakers of a language
            TotalPop += Co*kList[i] #the effect of having babies population of the 
world
            subsub[2] = int(subsub[2]) + int(1.5*cDict[subsub[1]]*(1-
Co)*LangT[LangN.index(rawList[i][1])]/
                                             LangT[j]*(TotalPop-
LangT[LangN.index(rawList[i][1])]))
            LangT[matrix.index(sub)] += int(1.5*cDict[subsub[1]]*(1-
Co)*LangT[LangN.index(rawList[i][1])]/
                                            LangT[j]*(TotalPop-
LangT[LangN.index(rawList[i][1])])) #the change of number of speakers of a language
            for i in range(len(sub)):
                try:
                    subsub[2] += (Co * (I(sub[i], subsub))) - (Co * (I(subsub, 
sub[i])))
                except:
                    continue
            for z in range(len(matrix)):
                for q in range(len(sub)):
                    try:
                        subsub[2] += (((1-
Co)*(I(matrix[z][q],subsub))*(matrix[z][q][2]/P(q)))
                                      - ((1-
Co)*(I(subsub,matrix[z][q]))*(matrix[z][sub.index[subsub]][2]/P(sub.index(subsub))))
)
                    except:
                        continue
    resultList1 = []
    resultList2 = []
    for j in range(len(matrix[0])):
        resultList1.append(matrix[0][j][2])
        resultList2.append(matrix[3][j][2])
    print("For English " + str(year + 1))
    for result1 in resultList1:
        print(result1)


print(matrix)
```