For office use only
T1 ________________
T2 ________________
T3 ________________
T4 ________________

Team Control Number
93434

Problem Chosen
B

For office use only
F1 ________________
F2 ________________
F3 ________________
F4 ________________

2018
MCM/ICM
Summary Sheet
(Your team's summary should be included as the first page of your electronic submission.)
Type a summary of your results on this page. Do not include the name of your school, advisor, or team members on this page.

Exploring Nonlinear Dynamics in Language over 100 years with the ”Frankenstein” Model

Summary

Each year, one can never be certain which problems will show up on the MCM, but the time constraints, difficultly, and creativity of three people is usually enough to forge an interesting model and derive some neat results. With Problem B, this year was no exception. In order to model language dynamics, we combined two established fields, language competition and disease modeling. With a little ingenuity and perhaps a bit of naivety, we created a **monster** of a coupled, compartmental differential equation model that we joking referred to as the Frankenstein of models, and the name stuck. Upon seeing our model's MATLAB function handle, in its not-so-glorious, explicit form, one might have guttural reaction similar to that in the quote below,

> *”If you had seen the man who thus capitulated for his safety, your surprise would have been boundless. His limbs were nearly frozen, and his body dreadfully emaciated by fatigue and suffering. I never saw a man in so wretched a condition.”*
>
> – Frankenstein by Mary Shelley

But through that reaction we charged steadily onward, hoping that our theory of using GDP as a good indicator of language conversion rate would prove correct. In the theory of demographics, GDP for each country has a large influence on the migratory patterns of populations because of its high correlation with countries of scientific and cultural prowess. One explicit example of this is brain drain, the emigration of highly trained or intelligent people from a particular country. In our model, we attempted to mimic this sense of movement due to the above factors by converting regional GDP data into GDP ”caused” by each language. We then normalized these GDP values for each language and created a transition matrix, $g_{L_i,Lj}$, inspired by our previous experience with disease models, that gave us a probability of a speakers likeliness to convert form language L_i to Language L_j. Because we did not know how large of an effect GDP should be, we multiplied the transition matrix by a scaling factor of γ. Through numerical analysis of our model, we found an optimal value of γ that minimized our model's error using 50 years of historical data (shown in Figures 1 and 2), collected from 50 different World Almanacs. With our found γ values we then predicted with moderate confidence that English will be the dominant language for total speakers in 10 years and Spanish will be dominant language for total speakers in 50 years. More concretely we can infer that GDP accounts for 28 percent of the language conversion in total speakers and 16.6 percent of the language conversion in native speakers.

Exploring Nonlinear Dynamics in Language over 100 years with the ”Frankenstein” Model

93434

February 13, 2018

1 Introduction

With people migrating, keeping long distance relationships, and companies growing and becoming international more frequent than in the past, language growth and decay has been of interest more recently.

Typically language growth is viewed as a chaotic system with many different factors - each ’competing’ with the other to become the majority. Because of the many factors, modeling languages with a basis on historical data is crucial to being able to pseudo-predict whether a language will grow or a decay in number of speakers throughout the future.

Looking at across historical data the top 20 Languages have been competing with each other over the past century with little change in those at the top 3. Interestingly, as time progresses there appears to be various languages approaching stable equilibrium between 0 and 100 million speakers, see Figure 1. Because of this we believe it is possible to model the language dynamics as a series of coupled differential equations.

1.1 Predominant Language Competition Models from the Literature

1.1.1 Abrams-Strogatz Model

The Abrams-Strogatz model is used to describe competition between two languages. The differential equation, shown below, models two languages X and Y ’competing’ for more speakers, denoted by x and y which are fractions of the population.

$$\frac{dx}{dt} = yP_{yx}(x, s_x) - xP(y, s_y) \tag{1}$$

The above equation finds the rate at which speakers of language X are changing. Since this is only a two variable system $\frac{dy}{dt}$ need not be modeled as it relates directly to $\frac{dx}{dt}$. This model makes a few assumptions:

1. The population size is constant and each person speaks

2. The population is highly connected, and the individuals interact at the same rate
3. The switch from one language to another is due to it's attractiveness
4. The attractiveness of a language increases with the number of speakers and it's status (s_x and s_y for X and Y respectively.)

1.1.2 Kandler Model

The Kandler Modeler is an extension of the Abrams-Strogatz model from equation (2) that introduces a bilingual group, denoted as 2, in addition to the original two monolingual groups, 1 and 3. The assumptions remain roughly the same, even though it adds a third compartment and a sense of space.

$$\frac{\delta u_1}{\delta t} = d_1 \Delta u_1 + a_1 u_1 (1 - \frac{u_1}{K - (u_2 - u_3)}) - c_{31} u_3 u_1 + c_{12} u_2 u_1 \tag{2}$$

$$\frac{\delta u_2}{\delta t} = d_2 \Delta u_2 + a_2 u_2 (1 - \frac{u_2}{K - (u_1 - u_3)}) + (c_{13} + c_{31}) u_1 u_3 - (c_{12} u_1 + c_{32} u_3) u_2 \tag{3}$$

$$\frac{\delta u_3}{\delta t} = d_3 \Delta u_3 + a_3 u_3 (1 - \frac{u_3}{K - (u_1 - u_2)}) - c_{13} u_1 u_3 + c_{32} u_2 u_3 \tag{4}$$

1.1.3 SIR Model

An SIR model is commonly used in disease modeling. It can be coupled shown in the general form below to capture a sense of movement between patches i and j where $i \neq j$.

$$S_i' = f(S_i) - \beta S_i I_i - \sum_{j \neq i} m(i,j) S_i + \sum_{j \neq i} m(j,i) S_j \tag{5}$$

$$I_i' = \beta S_i I_i - (v + b + \alpha) I_i - \sum_{j \neq i} m(i,j) I_i + \sum_{j \neq i} m(j,i) I_j \tag{6}$$

$$R_i' = v I_i - b R_i - \sum_{j \neq i} m(i,j) R_i + \sum_{j \neq i} m(j,i) R_j \tag{7}$$

The coupled, compartmental model, while simple, is a powerful tool that has been used and extended on for more that 40 years to describe the dynamics disease infection of people.

2 Definitions

- **Gross Domestic Product (GDP):** A monetary measure of the market value of all final goods and services produced in a year.
- **Rate of Natural Increase (RNI):** The crude birth rate minus the crude death rate.

- **Native Language:** A language that a person has been exposed to from birth.
- **Native Speakers of a Language:** The population of the world that has been exposed to a language from birth.
- **Total Speakers of a Language:** The population of the world that are able to speak a language at any point in time.
- **Transition Matrix:** A square matrix that gives the probability that a member of a population will stop speaking language i and start speaking language j, where the diagonal elements are 1.
- **Language Hub:** The major geographic area where languages are expected to be the most widely spoken. For example, Japanese and Chinese would be located in ’East Asia’.

Symbol	Meaning	Value
$P_{i,t}$	Population of speakers in language i at time t	Varies
L_i	Language i	Top 21 languages in 2017
Λ	Number of Languages	21
dt	Change in time in years	Varies
λ_i	Natural increase rate of language i	Based on regional data
Π_i	Number of regions	The 6 populated continents
γ	Scaling factor to adjust effect GDP has on P_i	Varies
g_{R_i}	GDP per region	Based on regional data
g_{L_i,R_j}	GDP for each language per region	Based on regional data
g_{L_i}	GDP for each language	Based on regional data
η_{L_i}	Population per language	Based on regional data
g_{norm}	Normalizing factor for GDP transition matrix	Based on regional data
$g_{L_i,Lj}$	GDP transition matrix	Based on regional data
sp_{R_i}	Amount of speakers per region	Based on regional data
sp_{L_i,R_j}	Amount of speakers for each language per region	Based on regional data

3 Assumptions

- **Distribution of speakers across regions are evenly distributed.** The even distribution allowed for the simplification of data cleaning and processing.
- **Proximity does not affect transfer.** The diffusion of speakers to different languages should not be affected by the proximity of those languages. The ideology behind this was that the internet and technology allow people to communicate over long distances without the need to learn other languages.
- **We also will use the following assumptions that were made originally in the Abrams-Strogatz model.**

- Languages do not innately change and thus remain steady as long as another language does not attract the speaker.
- Population is highly connected with no spatial structure.
- No distinction will be made between different uses of language in different social context.

- **Constant GDP and RNI and other factors were assumed for the top 21 languages.** Time did not permit to allow us to model the change in these factors into the future, which could have been done through modeling GDP of each language coupled with our model.
- **Languages can die out leading to other attractive languages to grow and replace.** This assumption is made in accordance to the Abrams-Strogatz model that was initially used for language extinction. When a language dies out, it cannot be 'revived'.
- **High-impact events with low probability were ignored.** This is per the problem statement.
- **The particular status of a language in a country determined how the countries would be labeled.** The countries of a certain language with status of 'Official', 'Co-Official', and 'Majority' were labeled as having 'native' speakers. Additionally, the status of 'Significant Minority' were labeled as the 'total' speakers.
- **GDP is a good indicator of speaker conversion rates.** Speakers will be more likely to switch languages if it means being apart of a group that is "better off" financially, industrially, or culturally. This has been observed in reality through what is commonly referred to as brain drain of a country.

4 A Model Built on Data

4.1 50 Years of Data

In order to check future placement, we looked 50 years into the past and analyzed the trends of the top 21 languages from 1966-2017 (Figure 1). That data was collected by going through 50 almanacs to obtain the language data. Data from 1966 to 2000 had the total number of speakers of that language; however, after 2000 the records were not updated every year and were changed to observe only native speakers. This was convenient, since the problem asked to analyze and predict for both total and native speakers of the top languages. This data was essential in performing model analysis and fitting parameters. Another set of data we deemed essential to have was the number of countries that spoke the languages of interest per region (Africa, Asia, Europe, North America, South

America, and Oceania). If we observed the countries themselves, the data would have been too overwhelming in such a short time and thus we considered only the countries per region. Additionally, the data collected for the factors were region-based. Before moving on to investigate the effect of these factors, the data from 1966 to 2017 for total language speakers and from 2000 to 2017 for native language speakers should be analyzed. (Note: for total language speakers the additional data point after 2000 was the 2017 data provided)

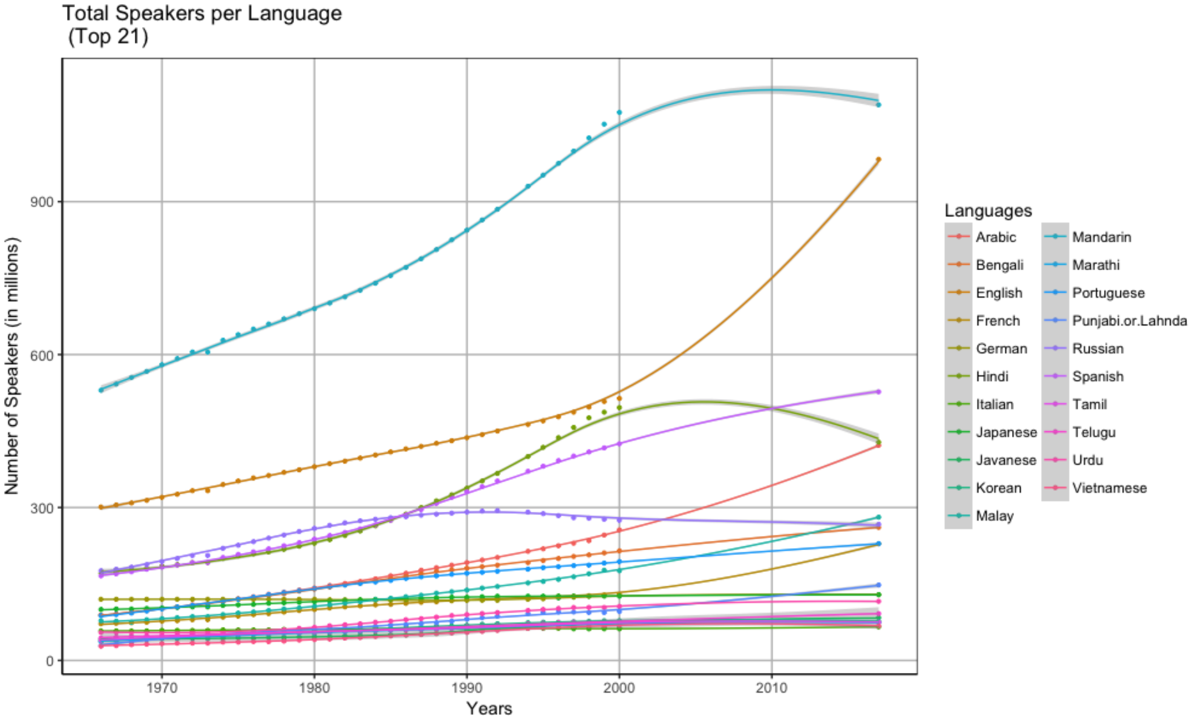


Figure 1: Top 21 Languages (Total Speakers) - 1966-2017

One of the first things to notice in Figure 1 is that some of the languages stay roughly stagnant. Slight increases in those languages near the bottom can be attributed to the general population increase of the world. Thus, a general statement can be made about languages that have a low total number of speakers. That is, they will remain stagnant and will not have a chance to replace the top ten languages. From the graph, the top ten languages are as follows: Mandarin, English, Spanish, Hindi, Arabic, Malay, Russian, Bengali, Portuguese, and French. Now we should observer the trends to estimate what will happen in the future, it is not possible to obtain actual data for these estimates - that will be the task of the model. We can also see that the top two languages dominate the charts, and that it is evident that English will overtake Mandarin in the foreseeable future. This is due to the decreasing nature of Mandarin and the seemingly exponential growth of English. It seems that Spanish and Arabic will be competing for the third spot; Hindi like Mandarin has a decreasing appearance as well. Arabic

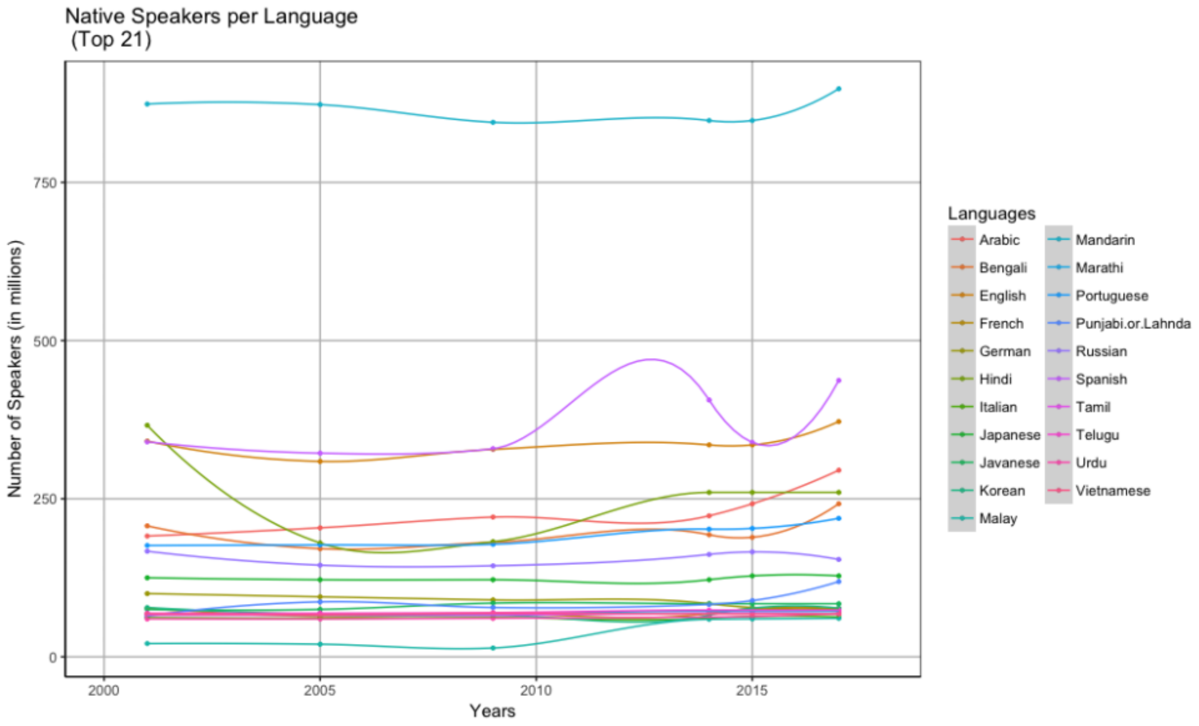


Figure 2: Top 21 Languages (Native Speakers) - 2001-2017

may eventually take over Spanish, due to its slight exponential nature, and then possibly overtake Mandarin in the far future. It seems the other languages in top ten are not really able to compete with the growth rates of the top five languages (Mandarin, English, Spanish, Hindi, Arabic). Thus based on historical data, the only evident chance in the foreseeable future for languages will occur among the top five languages. These are all important considerations when formulating a model.

In Figure 2, we have the change in number of native speakers over time at a mere six data points. From this, it is apparent that most of the languages stay constant with little to no growth or decay near 2017. Furthermore, we will ignore the hump between 2010 to 1025 in the data for Spanish, and just assume that was an outlier caused by poor line fit. From Figure 2, it seems clear that Mandarin will be the top language for native speakers, and that neither English nor any other language will replace it.It also seems the second most native spoken language is close tie between Spanish and English, with Spanish eventually winning out. Behind these two languages it seems that Arabic will catch up to them in the near future. Then Bengali, Hindi, and Portuguese fall shortly behind them, with Russian being below them. Bengali and Portuguese seems eventually outgrow Hindi as it seems to cap out. The remaining languages remain rather constant. By definition, the first language you learn as a child becomes your ’native’ language. It is a logical jump to assume a major factor in native speaker growth is rate of natural increase (the difference between crude birth rate and crude death rate) per language, which is most likely to remain roughly constant for developing countries. In general we have to remember that the span of the graph is only 17 years and is the most probable cause for the stagnation of native speakers for languages.

We can compare the trends between Native and Total speakers to really understand the language dynamics. For example, English natives seems stagnant but the total number of Exponential speakers is exponentially increasing as seen in Figure 1. This is most likely due to the fact that English is becoming a worldwide language and the amount of people learning English as a secondary language is increasing. Whereas, glancing at Mandarin it becomes clear that total number of speakers is falling, because it is not a global language and is most likely concentrated only in Asia. These and other factors have a huge affect on language dynamics, in this paper we go on to examine factors, such as GDP, population, and RNI. In order to extract useful information about these factors in regards to language change, data manipulation and transformations were needed to convert country and region statistics to language speaker statistics.

4.2 Data Transformations

To begin with, we note that the manipulation procedures for the 'Total Language Speakers' and 'Native Language Speakers' were the same but performed separately. For this reason, there will be no need to specify the type of language speakers for this portion of the paper. After obtaining the number of countries per region per language we were able to obtain a ratio. Which is the percentage that each region contributed to a language. Using the data of total number of speakers of the languages in 2017 we easily obtained the number of speakers of the specific language per region. Summing up regionally, we found the total number of speakers of the 21 languages for each region. Then we changed the factors per region to something more useful factors per language. The same methodology was used in this conversion for all the factors; Thus without loss of generality we will describe the calculation for GDP per region, g_{R_j}, to GDP per language, g_{L_i}. Through the use of the following equation we translated GDP of each region to GDP of each language per region. After this, we simply summed the GDP of each language per region over the six regions yielded the GDP per language. The following two equations displays this transformation.

$$g_{L_i,R_j} = \frac{(g_{R_j})(sp_{L_i,R_j})}{sp_{R_j}} \tag{8}$$

$$g_{L_i} = \sum_{j=1}^{\Pi} g_{L_i,R_j} \tag{9}$$

As mentioned earlier, this same formulation was done for the other two factors so that we could obtain the population per language, η_{L_i}, and the rate of natural increase per language, λ_{L_i}. The last and final step that was necessary to finish off parameterization of data was to obtain our transition matrix of the GDP, g_{L_i,L_j}, which was used to determine the flow or conversion of speakers from one language L_i to another L_j. In order to ensure that the transition matrix had values only between 0 and 1, a 'norm', g_{norm}, was introduced that took the ratio between the minimum and maximum GDP of the top 21 languages. The equation for obtaining the components of this matrix is as follows.

$$g_{norm} = \frac{\min\limits_{1 \leq m \leq \Lambda} g_{L_m}}{\max\limits_{1 \leq n \leq \Lambda} g_{L_n}} \tag{10}$$

$$g_{L_i,L_j} = \frac{g_{L_j}}{g_{L_i}} g_{norm} \tag{11}$$

$$g_{i,j} \epsilon [0,1], \quad i \neq j \tag{12}$$

Using this method, we transformed real world data into quantities for population, the natural rate of increase, and our GDP transition matrix. These will play an utmost important role in modelling language growth and decay throughout the future.

5 Formulation the Frankenstein Model

5.1 A Convenient Combination

Upon inspection of the Abram-Strogatz, Kandler, and SIR models from 1,2, and 5 respectively, we can see that are very similar. In fact, they are all types of compartmental ordinary differential equation models. The idea behind compartmental models is to capture the idea of movement or change of state in a population. Thus we had the idea to merge them all.

$$\frac{\delta P_i}{\delta t} = \lambda_i P_i - \sum_{j=1}^{\Lambda} \gamma g_{i,j} P_i + \sum_{j=1}^{\Lambda} \gamma g_{j,i} P_j \tag{13}$$

Our model includes to include birth and death rates in the form of the RNI, λ, and changes dynamically as time changes and speakers decide to move to different languages over time as they are influenced by dominant

GDP's, $g_{i,j}$. In order to to determine the appropriate effect that GDP has on language conversion rates, we also included a scaling term, γ. We discuss the effect and possible physical value of gamma later in the paper.

An essential component of our model is the transition matrix, denoted by $g_{i,j}$. This matrix details the conversion rates of speaker from language i to language j where $i \neq j$. As the populations increase the effect of this matrix will also increase causing larger and larger conversions to or from certain languages. As the populations decrease the conversion will have less effect on overall dynamics of the model.

Perhaps the most intriguing aspect of the model is the simultaneous competition of 21 languages. Even though the transition matrix and RNI are held constant because of the limited amount of time to collect historical data, each language will be directly affected by the ”strength”- the amount of speakers in each language, and ”prominence”- the ratio of each languages GDP with respect to each other.

5.2 A Compact Generalization, an Ugly Expansion, a Decent Name

While the general form of our model, Equation 13, is compact and easier to understand, the fully defined set of coupled differential equations is a monstrous beast that we joking referred to as the Frankenstein of models. After coding up the model in MATLAB our attitude toward the function handle of our coupled differential equations paralleled the tone of the quote below.

> *”If you had seen the man who thus capitulated for his safety, your surprise would have been boundless. His limbs were nearly frozen, and his body dreadfully emaciated by fatigue and suffering. I never saw a man in so wretched a condition.”*
>
> – Frankenstein by Mary Shelley

Explicitly typed out for all 21 languages we analyzed, the aptly named Frankenstein model contains exactly 16435 (characters without spaces which equates over 900 individual terms that are the product of between two and three variables or coefficients.

6 Analysis of the Model using Historical Data

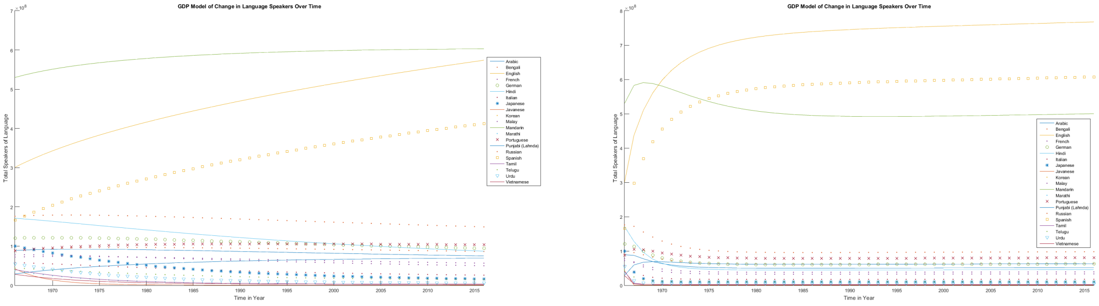
6.1 Numerical Integration of Coupled Differential Equations

We used the **ode45** function in MATLAB to numerically integrate our system of 21 coupled differential equations, where each time step represented a year. With this methodology, we ran two sets of simulations with our model.The first was of total speakers for each language where the time domain started in 1966, the first data point in Figure 1.

The second was of native speakers for each language where the time domain started in 2001, the first data point in Figure 2. We should also note that the GDP transition matrix from Equation 11 was reevaluated based on the total and native speakers respectively and that the initial amount speakers was adjusted to reflected the historical data.

We then plotted all 21 solutions to our Frankenstein model and labeled each language according to its particular GDP, RNI, and number of speakers at our first historical data point. This gave us multiples plots for both the total speaker and the native speakers that we later compare to the historical data we collected earlier. We hope this process of comparing our model to historical data gives us robustness and accuracy in our eventual predictions 50 years into the future.

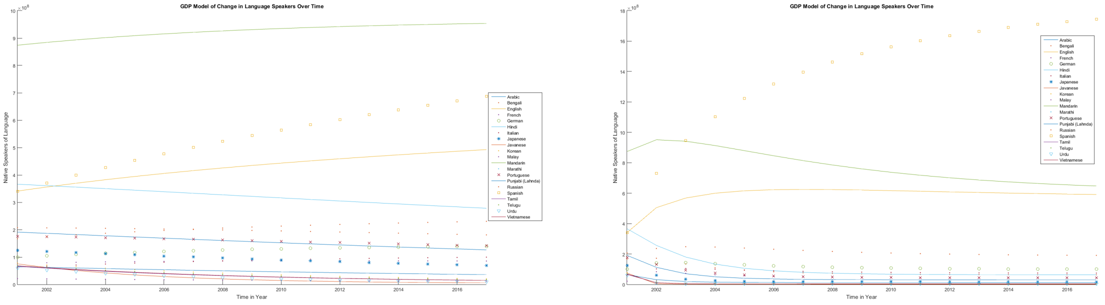
To gain a further understanding of the dynamics of our model, we experimentally altered the γ in order to match trends over the last 50 years to pinpoint the effect of GDP on language conversion. The effect of different γ values with two different transition matrices and initial value points can be seen in Figure 3 and 4.



(a) Results of the model for total speakers with $\gamma = .05$

(b) Results of the model for total speakers with $\gamma = 1$

Figure 3: Numerical plots for total speakers with varying γ



(a) Results of the model for native speakers with $\gamma = .05$

(b) Results of the model for native speakers with $\gamma = 1$

Figure 4: Numerical plots for native speakers with varying γ

Despite the assumptions we made for simplicity and despite our model's narrow focus on birth rates and GDP we can note some similarities in trends in the Frankenstein model and trends in the historical data. Namely, these trends include:

- Trend 1: According to Figure 3a with $\gamma = .05$ we concluded that Hindi and Bengali declined periodically while English, Spanish, and Mandarin grew logarithmically at about the same rate. This leads to the top 5 languages in 2017 being Mandarin, English, Spanish, Bengali, and Portuguese. The most accurate of the languages appears to be Mandarin and English which seem to be converging and potentially swapping ranks in the near future. This is the closest to matching our historical data, however the other languages lower in accuracy over time. This could imply that English and Mandarin are affected by about 5 percent of the GDP.
- Trend 2: We also noticed that Figure 3b with $\gamma = 1$ seemed to most accurately account for the short term rate of increase according to historical data. Although Mandarin and English swap at a much earlier date the sharp increase is not accounted for as accurately in the other models. This could perhaps imply that GDP affects language growth short term.

6.2 Optimization of an Unknown Parameter

6.2.1 Creating a Cost Function

In order to obtain an approximation for γ, we calculated the error using Equation 14 for each of the t years we had historical data points for.

$$Error = \max\left(H_{L_i,t} - P_{i,t}\right)^2 \qquad i\epsilon[1, 2, ..., 21] \tag{14}$$

Using this function, we minimized the error of our model with respect to γ through 1966 to 2017 using the **fminsearch** function in MATLAB. This gave us two distinct ”optimal” gamma for the native speaker and total speaker models.

$$\gamma_{total} = .2804 \tag{15}$$

$$\gamma_{native} = .1668 \tag{16}$$

Plots of our model with these optimal γ values are located in Figures 5 and 6 for total and native speaker simulations respectively. We discuss the dynamics and implications that these γ values introduce later in the paper.

6.2.2 A Not-So-Optimal γ

After optimizing our model over γ to reduce the cost function in Equation 14, we expected to see our model plot closer to the plots of historical data in Figures 1 and 2. Unfortunately, this was not the case, as seen in Figures 5 and 6. By comparing Figures 3a and 15 and Figures 4a and 16, it is easy to gather that .05 is seemingly a better value than the optimal γ when looking at corresponding trends in from Figures 1 and 2. Although the plots using the optimal γ values do not exactly follow the trends of the historical, we can still derive real world analogues to the found values. One of the first properties to notice is the smaller γ is the less drastic the rate of growth and decay is.

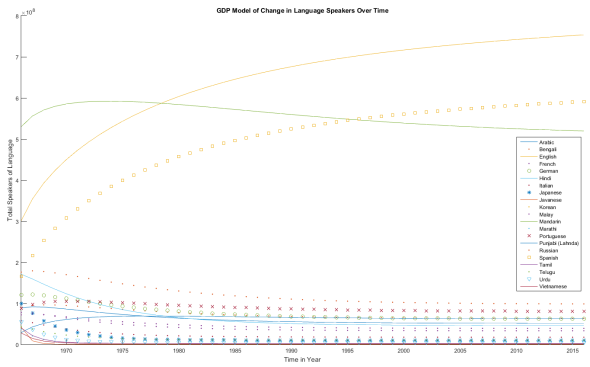


Figure 5: Results of the model for total speakers with optimal $\gamma = .2804$

Another interesting interpretation of the γ values is the importance of GDP during the decision to switch languages. Under this interpretation, approximately 16.68 percent of native speakers are influenced by GDP, and approximately 28.04 percent of total speaker are influenced by GDP. These values could be used as a scaling conversion coefficient in future models that use GDP as a factor to predict language dynamics.

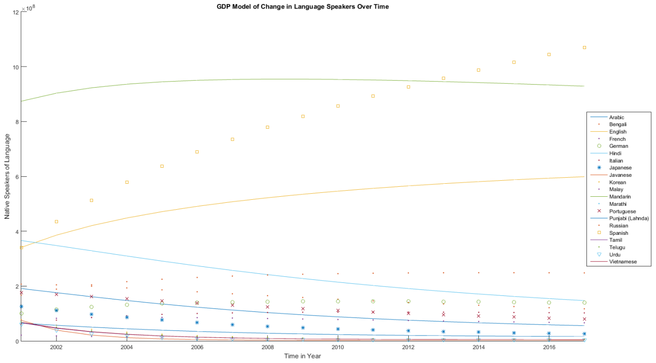


Figure 6: Results of the model for total speakers with optimal $\gamma = .1668$

7 Migration Patterns can lead to Language Growth and Decay

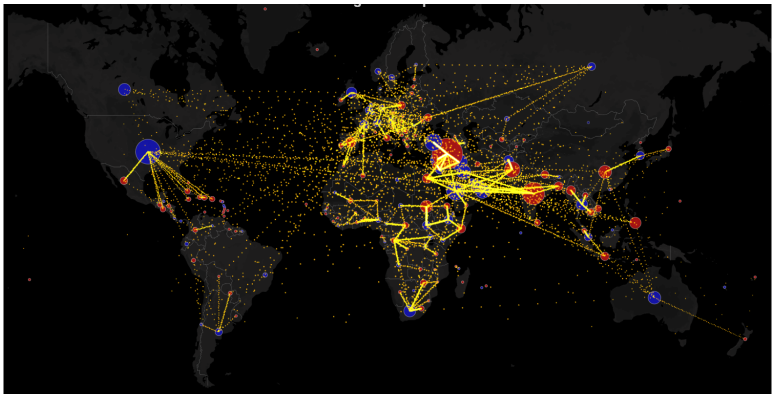


Figure 7: Human Migratory Patterns from 2010 - 2015

7.1 Assumptions in Migration Patterns

Before analyzing if and how human migration patterns could effect geographic distributions of the top 21 languages, it is important to discuss Figure 7. The map shows net migration, the difference between immigration and emigration. Blue regions indicate positive net migration (higher inflows), whereas red regions indicate negative net migration. Furthermore the map was interactive so clicking on a region would show migration flows in and out of that region. This helped the analysis of human migration patterns and the predicted effect it would have on the top 21 languages.

A couple of assumptions had to be made for human migration and its effect on estimated language movement.

- **Human migration patterns for the next 50 years will remain constant.** The data from this interactive map only displayed from years 2010 to 2015, prediction data for the next 50 years was too difficult to find in a format close to this map.
- **Net immigration heavily influences total number of speakers of a language.** Those that migrate to a new region will need to learn their new home's native and majority language to effectively communicate in their environment. Thus, if there is large positive net migration to some region then the number of total speakers of that region's language will increase.
- **Net immigration will have an effect on native number of speakers of a language, but will take a longer time to develop.** Those that have migrated to a new region will eventually reproduce children native to

that region. It is safe to assume that child's native language will be of that region; therefore, the number of speakers of the native language of that region will inadvertently increase.

7.2 Migration Patterns in Language Hubs

In order to estimate predicted trends of geographically language movement, we first had to give major regions a dominant language, so that those areas would become language hubs. The labelling of dominant languages in a region was achieved by observing the ratio of total countries in the region to the total countries that spoke that language. Sometimes the majority ratio for a language was not extremely clear (the second/third largest ratio would very close to the first), in this case the language was appropriated to be dominant in multiple regions. Note that for the sake of labelling all languages, a region can have multiple dominant languages, but in general, languages in the top ten will eventually dominant their respective regions and maybe other regions depending on migration patterns. Larger regions, like Asia and Africa were split into smaller subregions which are listed as follows - Asia (South, North, East, South East, Middle East) and Africa (South, North, and Central). Therefore, we can examine immigration patterns in these regions and make inferences about the future movement of languages. It is important to note, if the region has a positive net migration than the number of both total and native language speakers increases. The impact of migration on total language speakers is larger and more immediate than native speakers. The opposite occurs for a region with negative net migration. Below we examine language movement per region, then we summarize overall language growth of the top 10 languages (the remaining according to the data and model will stay stagnant regardless).

- **Asia - Middle East:** In this region the only dominant language is Arabic. There is a large negative migration bubble in the Middle East, which is surrounded by smaller but significant positive migration regions. Overall, there seems to be a somewhat positive increase of Arabic speakers in this subregion.

- **Asia - South:** This region is mainly the Indian subcontinent and contains many of the Indo-Aryan languages (Bengali, Hindi, Punjabi, Tamil, Telugu, Marathi, and Urdu) out of these Hindi dominates them by a considerable amount. However, in India there is a large negative net migration. This shows the decrease in Hindi speakers in the general since this subregion is the only hub for Hindi.

- **Asia - East:** This region includes the languages of Mandarin and Japanese. We can individually see the trends in both, with the assumption that China is the Mandarin hub and Japan is the Japanese hub. China has a net outflow of speakers to other parts of the world, mainly to North America. Thus Mandarin is losing speakers to English. In a similar fashion, Japanese is losing speakers as well to many areas of the world. For both of these languages, this subregion is their only hub.

- **Asia - South East:** This region includes countries such as the Koreas, Vietnam, and Malayasia. Malay will dominant this region. There isn't much net migration and much of it occurs between neighboring countries. Thus, the prediction here is that these languages will remain mostly constant.
- **Africa - Central:** This subregion is predominantly French. There is little movement here, if any, most of it stays within its subregion; therefore, we can safely predict no net increase in French speakers due to this subregion.
- **Africa - North:** This subregion is predominantly Arabic. There is little movement here, if any, most of it stays within its subregion; therefore, we can safely predict no net increase in Arabic speakers due to this subregion.
- **Africa - South:** This subregion is predominantly English. There is little movement here, if any, most of it stays within its subregion; therefore, we can safely predict no net increase in English speakers due to this subregion.
- **Oceania:** This region is predominantly dominated by Javanese. There isn't much migration globally, but there is notable lost of Javanese speakers to Australia where the main language is English. Thus there is small net gain in English speakers.
- **Europe:** There are multiple major Indo-European languages in this region, which are English, French, German, Italian, Portuguese, and Spanish. Again, it seems like the majority of the migration occurs within Europe. There seems to be net increases in Germany, France, and the United Kingdom whereas a negative migration in Spain, Portugal and Italy. The greatest net increase was in the United Kingdom, thus there was a decent increase in English speakers in comparison to other language speakers. Even though it seems that Spain had a loss of Spanish speakers, those speakers went to other Spanish-speaking hubs; therefore, the number of Spanish speakers due to Europe stayed constant. A similar but on a smaller scale case occurred in Portugal. In general, there was an increase of English speakers in this region.

- **North America:** This region can split into smaller subregions. Canada can be considered the French hub, United States can be considered the English hub, and Central America can be considered the Spanish hub. Starting with Canada, there is a decent-sized positive net migration from places all over the world. Therefore there is a net gain in French speakers in this region, some of the largest migration to Canada seems to be from the Hindi hub and the Mandarin hub. Similar to Canada but with much larger increase of positive migration, the English hub has a large increase. It is important to note that a large majority of speakers from the Spanish hub are migrating to the English 'Hub. Thus, there is an increase of French speakers, a significant increase of English speakers, and a decline of Spanish speakers in this region.

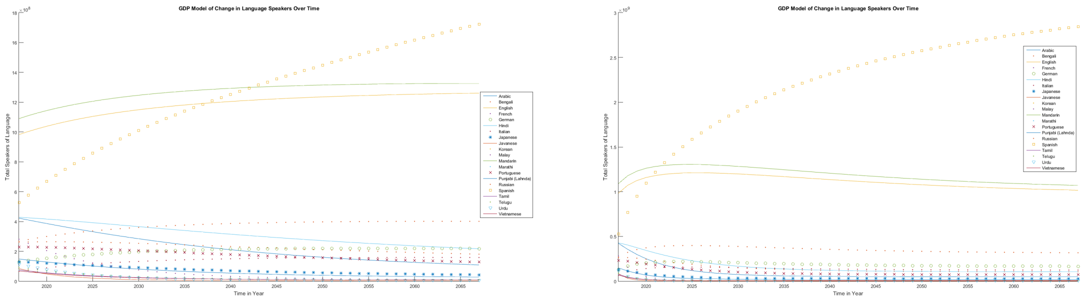
- **South America:** In this region, the two major languages are Spanish and Portuguese. On average there does not seem to be much migration from regions outside of South America. The little migration that does happen shows a small increase in both Spanish and Portuguese speakers.

7.3 Summary: Migration Patterns affects Language Dynamics

In summary, Figure 7 provided interesting insights to the language dynamics per region. In order to understand global language dynamics, we see the net increase of speakers for languages over all the regions. Do take note, that the reason for these migration patterns is not within the scope of this paper and will not be discussed. In general, there was an increase in Arabic speakers due to migration from other subregions in the Asia region. This increase is in accordance to the model. Hindi along with other Indo-Aryan languages that we tracked saw a general decline of speakers which is in accordance to the model trends. Japanese, and more importantly, Mandarin both saw a decline of speakers. This concurs with the model and the net migration as discussed earlier. From the model, we see an exponential growth of English speakers this is in agreement to the large increase of English speakers due to migration. The migration to many of the English hubs was always positive and usually came form many regions throughout the world. Which in turn would generally lead to a decrease of language hub of that region or subregion and an increase of English speakers. French and Portuguese saw a small net positive migration increase. In general most of the migration data concurred with our model. However, the one language that did not have an agreement between the model and migration data was Spanish. According to migration data, Spanish speakers should be in somewhat of a decline due to the overall slight negative net migration; however, according to the model and data Spanish is growing. This may be due to factors such as rate of natural increase and others that we have not accounted for, but migration patterns do have an effect. Observing other language growth, such as English, which appear to be exponential, Spanish does not have that same trend instead it has linear growth. We hypothesize that the cause of this is due to the negative effect that migration had on Spanish, it is like a dampening effect on its growth.

So, with the assumption that migration data will stay the same for the next 50 years, then net migration will have an impact on language growth. Positive net migration will act as a positive effect on its growth and increase the number of speakers in both the total and native category. As discussed earlier, it will have a larger impact on total number of speakers per language than native. At the same time, a negative net migration can dampen language growth as seen in the Spanish speakers trend line.

8 Results: Predictions for 50 Years into the Future



(a) Results of the model for total speakers with $\gamma = .05$

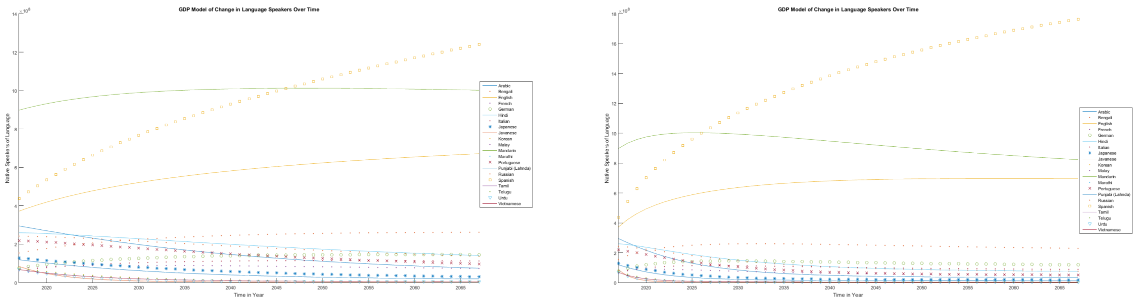
(b) Results of the model for total speakers with $\gamma = .2804$

Figure 8: Numerical plots for prediction of amount total speakers with varying γ

Using these Figure 8 we can analyze the effects of γ on the model. The optimal γ value is 0.2804 and the γ for the other graph displayed is 0.05 which is considerably smaller than the optimal γ. This optima γ was found by performing an error analysis of our model's prediction from 1966 to 2017 against the data collected, the same was done for the native speakers.

In both figures the top three languages do not change whatsoever. They are Spanish, Mandarin, and English, respectively. Spanish seems to become the most widely spoken language in both cases and Mandarin is larger than English. Even though, the trends from the original data suggest otherwise. Specifically, that English seems to be the language of the future. This may be due to factors that have not been taken into consideration in our model. Also regardless of the γ value, the other languages seem to eventually die out. This was expected as per our assumptions. The exception to this was Russian, which seems to reach an equilibrium number of speakers. In general the trends of the data do not change too much; however, the scale and the rate of change is vastly different between the two γ values.

At lower γ values the trends change much slower and are more gradual and the scale of the number of speakers is smaller as well. For example, the number of speakers of Spanish in 2066 for $\gamma = 0.05$ approaches 1.8 billion whereas for the optimal γ the number approaches 3 billion. Take not that at the optimal γ Spanish overtakes English and Mandarin much earlier than at lower γ values. When $\gamma = 0.05$, Spanish overtook Mandarin around the year 2043, whereas at the optimal γ, Spanish overtook Mandarin around the year 2022. Also, take note that the bottom languages die out much more slowly (some do not even seem to die out) when the γ is a smaller value. In a sense, lower γ values seem to lower language dynamics in the sense that languages grow and die out more slowly.



(a) Results of the model for native speakers with $\gamma = .05$ (b) Results of the model for native speakers with $\gamma = .1668$

Figure 9: Numerical plots for prediction of amount native speakers with varying γ

A similar analysis can be done using the graphs made for the amount of native speakers.

8.0.1 Recommendation Short Term - 10 years

Over the short term we believe that the following will be the top 3 languages for total speakers: English, Mandarin, Spanish. Over the short term we believe that the following will be the top 3 languages for native speakers: Mandarin, Spanish, English. Additionally we also believe that over 10 years at least 9 out of the top 10 languages in the top ten will remain the same.

8.0.2 Recommendation Long Term - 50 years

Over the long term we believe that the following will be the top 3 languages for total speakers: Spanish, English, Mandarin. Over the long term we believe that the following will be the top 3 languages for native speakers: Spanish, Mandarin, English. Additionally we also believe that over 50 years at least 8 out of the top 10 languages in the top ten will remain the same.

8.1 A Constantly Changing World (Future Work)

Though our model was able to accurately fit the obtained data points and predict into the future, there could have been many improvements made from the data collection to the model and parameters. Here is a list for a few suggestions that could be made:

- **Parameterization of distance and migration data.** The parameters that we believed to have the largest effect on language dynamics were the GDP, population, rate of natural increase, distance and migration. If time had permit, the idea would have been to implement distance and migration as parameters for language data. Early in the paper, it was observed that migration and in turn distance could have an effect on language growth. We were extremely close to fulfilling the distance parameterization and had even completed obtaining the data and transformation of it. That is, a 21x6 matrix of language contribution per region and a 6x6 matrix the distance between each regions geographic center were calculated. However, it would have been better to use subregions as examined in the migration section of the paper. Furthermore, the parameterization of migration per language could have been put to use as well. Increasing the number of parameters used in the model from three to five, would have definitely increased the accuracy of our model.

- **A complete data in language speakers from 1966 to 2017 and parameters of interest** Our data was quite strong for the total number of speakers from 1966 to 2017, since we manually went through old almanacs to collect this data. However, the data for native speakers was nowhere as thorough just because of the lack of data that could be obtained freely. Furthermore,the parameters that we used were set at a fixed year, for example we used the GDP from 2016 as the parameter throughout 50 years of data. In reality, these parameters are actually time dependent which should be taken into consideration. On top of that, the parameters were obtained per region; however, if parameters per subregion (the Language hubs) were used then it would have been another way to strengthen our model and accurately predict language dynamics into the future.

- **Perform nonlinear regression on data collected to compare to the model's prediction of 50 years into the future** We optimized the model to data collected from 1966 to 2017 and then predicted 50 years into the future with that model after optimizing it. A suggestion to check the accuracy of the model for 50 years into future would be to take our data and perform a nonlinear regression. Then we could calculate the error of the model's prediction 50 years into the future and the nonlinear regression's prediction 50 years into the future. This would help us obtain an accurate measure of how well our model performs against well-known nonlinear regression models and could further optimize our parameters.

9 Letter

To the Chief Operating Officer:

We compiled data from 1966-2017 for the number of speakers for each the current top 21 languages from 1966-2017. We used R to interpret and graph the data in order to gain better insight into how the languages have grown over time. We analyzed the top 21 languages to give us a larger scope in the event that a particular language grows or drops within the top 5 languages. Our model can predict the future number of native and total speakers. As per your request we have created a model and analyzed trends of language growth over the next 50 years. I believe you will be pleased with our results.

We were able to spot trends over the past 50 years and compare these trends with our model's output.The following is a recommendation based upon our model's predictions and should be considered when making the transition internationally. We know that language competition is a big consideration international companies need to accommodate for when looking at long time commitments such as relocation and creating new offices, and so we have taken the utmost care to create an accurate model that accounts for data such as GDP, RNI, and Language Interaction distribution. According to our findings we would like to recommend that you consider GDP when factoring in where to send your employees. Through numerical analysis of our model, we noticed that it would be better to send your employee's overseas and have them speak one of these 5 languages: English, Mandarin, Spanish, or Arabic.

We also recommend you take in to account short term versus long term contracts as the top languages do fluctuate around the 10 year mark. Over the short term (¡10 years) we believe that the following will be the top 3 languages in order of rank for total speakers: English, Mandarin, Spanish and that the following will be the top 3 in ranked order for native speakers: Mandarin, Spanish, English. Additionally we also believe that over 10 years at least 9 out of the top 10 languages in the top ten will remain the same. Over the long term there were very minute alterations within the top languages with only small ordering differences between the dominant languages. We predicted that for long term the leading Total speakers are Spanish, English, Mandarin and the leaders for native speakers are Spanish, Mandarin, English. Additionally we also believe that over 50 years at least 8 out of the top 10 languages in the top ten will remain the same.

Every two weeks, another language dies. With this model we believe you will be in better hands while calling the shots, and that it will be be the most beneficial to restrict yourselves to the top 3 languages spoken.

Sincerely,

Team 93434

References

1. Amano T. (2014) Global distribution and drivers of language extinction risk

   http://rspb.royalsocietypublishing.org/content/281/1793/20141574.figures-only

2. Isern N., Fort J. (2014) Language extinction and linguistic fronts

   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3973370/

3. Kandler A., (2009) Demography and Language Competition

   https://www.jstor.org/stable/41466598?seq=6page$_{s}can_{t}ab_{c}ontents$

4. Liu Y. (2009) Modeling Language Competition

   guava.physics.uiuc.edu/ nigel/courses/569/Essays$_{F}all2009/files/liu$*.pdf*

5. Young H. (2013) The Digital Language Divide

   http://labs.theguardian.com/digital-language-divide/

6. Nicholson C. (2018) 12 Interesting Facts about Languages

   http://lingualinx.com/blog/12-interesting-facts-languages/

7. Abrams D. , Strogatz S. (2003) Linguistics: Modelling the dynamics of language death

   https://www.nature.com/articles/424900a

8. Patriarca M. , Leppanen T. (2004) Statistical Mechanics and its Applications

   https://www.sciencedirect.com/science/article/pii/S0378437104002511

9. Fujie R. , Aihara K., Masuda N. (2012) A Model of Competition Among More than Two Languages

   https://link.springer.com/article/10.1007/s10955-012-0613-8

10. Simons, Gary F. and Charles D. Fennig (eds.). 2017. Ethnologue: Languages of the World, Twentieth edition. Dallas, Texas: SIL International. Online version: *http://www.ethnologue.com*

11. Wikipedia ”Laplacian Matrix” (Last updated 2018)

    https://en.wikipedia.org/wiki/Laplacian$_{m}atrix$

12. Wikipedia ”List of Countries by spoken languages” (Last updated 2018)

    https://en.wikipedia.org/wiki/List$_{o}f_{c}ountries_{b}y_{s}poken_{l}anguagesFrench$

13. Wikipedia ”List ofcontinents by population” (Last updated 2018)

    https://en.wikipedia.org/wiki/List$_{o}f_{c}ontinents_{b}y_{p}opulation$

14. Department of Economic and Social Affairs, (2009) World Mortality Report 2009

    http://www.un.org/esa/population/publications/worldmortality/WMR2009$_{R}eport_{f}inal.pdf$

15. Department of Economic and Social Affairs, (2017) World Mortality 2017 *http://www.un.org/en/development/desa/population/publications/pdf/mortality/World-Mortality-2017-Data-Booklet.pdf*

16. Wikipedia ”Rate of Natural Increase” (Last updated 2018)

    https://en.wikipedia.org/wiki/Rate$_{o}f_{n}atural_{i}ncrease$

17. Abrams D. , Strogatz S. (2003) Modelling the dynamics of Language Death ()

    https://www.math.uh.edu/ zpkilpat/teaching/math4309/project/nature03$_{a}brams.pdf$

18. Wallace D., Hu-Wang E. , Chen M.Patch (2012) Sir Models on the K-regular Graph

    www.math.dartmouth.edu/ dwallace/papers/WallaceChenHu.pdf

19. Wikipedia ”First Language” (Last updated 2018)

    https://en.wikipedia.org/wiki/First$_{l}anguage$

20. M. Anderson and R. M. and May, Nature 280, 361 (1979)

21. Wikipedia ”Human Capital Flight” (Last Updated 2018)

    https://en.wikipedia.org/wiki/Human$_{c}apital_{f}light$

22. Wikipedia ”Gross Domestic Product” (Last updated 2018)

    https://en.wikipedia.org/wiki/Gross$_{d}omestic_{p}roduct$

23. Figure 9 of Human Migration Patterns from 2010 - 2015 Pub. (2016)

    http://metrocosm.com/global-immigration-map/

24. Shelley, Mary. *”Frankenstein”* Lackington, Hughes, Harding, Mavor Jones (1818)