For office use only	Team Control Number	For office use only
T1 ________________	111111	F1 ________________
T2 ________________		F2 ________________
T3 ________________	Problem Chosen	F3 ________________
T4 ________________	ABCDEF	F4 ________________

Executive Summary

Companies must be able to determine how to expand their endeavors to one day become international. When planning where to establish a new office, many factors come into play. One of these factors for a large multinational service company is the languages spoken in the target country. Properly targeting the correct language demographic is an important consideration in deciding where to make large investments.

The issue of predicting spoken languages around the world is not only applicable to a single business. Any international business should be interested in the languages that will be spoken by their employees and customers in a target country to best provide quality products and services. Having predictions for language distributions could also prove useful to other entities such as nations wishing to base policy decisions on the cultural and language profiles of their citizens.

Not much work has been in the area of predicting language distributions over time. Thus, we have found the following:

English retains the most widely spoken language. Today, English is spoken by more people than any other language, approximately 1.2 billion. Fifty years from now we predict the population of English speakers to reach 1.45 billion followed closely by Mandarin Chinese at 1.28 billion.

The top ten most spoken languages should remain at the top for the foreseeable future. While the order of the largest languages changes slightly over the simulated period, the same languages appear in the list.

Countries will become more diverse on average with respect to their language distribution. Due to the effect of increased emigration and modern communication techniques, we predict that countries will shift from being relatively monolingual to a state of polylingualism. In extreme cases, we even see countries abandoning their native language as globalization continues.

In regard to current trends, we suggest that a company currently stationed in New York City, United States and Shanghai, China to expand by opening offices in:

- India – speaking Hindustani and Bengali
- Brazil – speaking Portuguese and Spanish
- Russia – speaking Russian and Mandarin
- France – speaking French and Spanish
- Japan – speaking German and Mandarin

Contents

1 - Introduction

When deciding on a location for an additional international branch of a company, linguistic factors play an immense role on the effectiveness of that branch. Whether or not your new addition contributes to the business' international identity is crucial in determining whether or it was worth the company's resources. Not only is it difficult to decide where to expand today, but even more so considering the next 50 years. If valuable company time and money are spent building a branch that is then deemed non-optimal, or even useless, this could pose an enormous long-term issue surrounding that branch.

In addressing this concern, many complications surrounding the lack of consistent data arose, many of which are most likely stemmed in the difficulty of collecting data from "enough" languages to accurately represent the global trends. As a result of this, there are few, if any, models that predict the population of a given language, let alone models that can use this to help justify the construction of an additional branch in a given location. Nevertheless, we have developed and suggested an implementation using a computer simulation that takes advantage of the relationship between geographical population growth and current language data to attempt to predict the number of speakers and their location to best advise this service company on where to continue their growth.

2 - Assumptions

We assume that population growth discretely recedes until 2055 when the United Nations predicts we reach carrying capacity circa 11.4 Billion people.

We assume that a total national population fluency of a given language that fell below 100,000 people was ignored in collecting data. In an attempt to only investigate realistic locations for an international business, minute numbers were ignored in amassing linguistic rates.

We assume that the language that an immigrant's child speaks is equally proportional to the distribution of languages spoken in that country. Since data on Immigration rates is not widely available, and there remains a grey area regarding which language is native for an immigrant's child, we made the assumption that the children in question were more exclusively swayed by the country in which they are raised, rather than the parents' native language.

We exclusively tracked the 26 most spoken languages in 2017 excluding 3 languages due to extreme lack of consistent data. Out of the 3 main sources we collected data from, none included sufficient data about Malay, Persian, and Hausa, and left us concerned about the accuracy of the model if these languages were included. Thus we excluded them entirely.

In addition, we assume that emigration from a country is constantly proportional to the population of that country. Emigration data is much harder to find than that of Immigration so unfortunately, we were forced to assume that Emigration was constant, and would later test the sensitivity of this assumption later in the process.

We also assume emigrants from a region disperse roughly uniformly to other countries.

We also assume that Social, Government, Technology, and Business Influence is proportional to population. We assumed this not only because of the raw lack of data, but also because we believe that over time, the influences for and against all languages will sum to 0 and thus be generally excluded from the model. Since we recognize that this may be unlikely, we also tested this in our sensitivity analysis.

3 - Definitions

L1 - A person's native language

L2 - An additional language a person may speak

Country - A country recognized by the United Nations

4 - Methods

Using these assumptions, we created a deterministic model to find the language growth not only for the globe, but for any given country. This was implemented using Python and allowed for data sets to be introduced to the system from Excel (xlsx) files. We iterated through each year, using a limited growth differential equation, to discover the future for the 23 languages we are investigating. In addition, we accounted for immigration, emigration, and any influence factors that may sway language rates within this model to best represent each language.

4.1 - Deterministic Model of Geographic Language Distributions

Let the change in population of speakers for a given language ℓ in a given country c be represented by:

$$\frac{dN}{dt} = \frac{New\ native}{speakers\ born} - \frac{Emigrating}{Population} + \frac{Immigrating}{Population} + \frac{Change\ in\ L2\ due\ to}{various\ influences}$$

$$\frac{dN}{dt} = \left[rN_c \frac{k - N_c}{k} \rho_{c\ell}\right] - [\rho_{c\ell} e_n N_c] + \sum_{\mathcal{M}-\{c\}} [\rho_{m\ell} e_m N_m i_m^c] + [I_\ell (1 - \rho_{c\ell}) N_c]$$

ℓ:= language in question	$\rho_{x\ell}$:= distribution of language ℓ in country x
c:= country in question	e_x:= emigration rate of country x
N_x:= current population of x	i_j^k:= immigration rate of country j to country k
r:= rate of growth of c	I_ℓ:= net influence rate for ℓ
k:= carrying capacity of c	$\mathcal{M}$:= Set of all countries

This model is applied to each of the 23 languages in all 169 countries that we are concerned with for a total of 3,887 populations being tracked. These are then compiled to provide insight into the populations and distributions of those languages in the future.

Because I_ℓ and $\rho_{x\ell}$ change every year, this model cannot be represented accurately as a continuous model. Therefore, the same model represented discretely follows:

$$\Delta N = \left[\left[rN_c \frac{k - N_c}{k} \rho_{c\ell}\right] - [\rho_{c\ell} e_n N_c] + \sum_{\mathcal{M}-\{c\}} [\rho_{m\ell} e_m N_m i_m^c] + [I_\ell (1 - \rho_{c\ell}) N_c]\right] \Delta t$$

Where, Δt is one year.

4.1.1 - New Native Speakers

To handle the influx of native speakers of various languages we utilized a limited population growth model. For this we considered the intrinsic growth rate r, the carrying capacity k, and the local language distribution $\rho_{x\ell}$. This language distribution tracks the proportion of individuals in a country who speak the language, either natively or as a secondary speaker.

This portion of the model predicts the number of new individuals that will be born in the country in question and then determines the number of those individuals that will speak each language.

4.1.2 - Migration

A major component in considering the geographic distribution of languages over time is the effect of migration on local language speaking populations. To model emigration we assumed that the number of people who leave a country is constantly proportional to the population of that country. For example, India would have a larger population of emigrants than Cambodia. This is represented by e_x. To handle the immigration to a country from all other countries, we described the number of individuals migrating from one country to another as a fraction of the number of people emigrating from the home country.

This section of the model behaves similarly to a diffusion reaction model. As time progresses, it is expected to reach an equilibrium point where all countries have equal proportions of each language.

4.1.3 - Internal Influences

Various internal influences play a factor in the number of people in a country who learn a language in a given year. For our purposes, we considered four different primary factors: government, business and international commerce, social pressures, and technological suppression. These all provide varying positive and negative influences. It can also be assumed that these influences change over time, as government policies change or centers of commerce shift.

4.2 - Development

Before landing on a population and migration model, we considered basing the spread of language on a model for the spread of an infectious disease, such as HIV. We ultimately decided against this approach because of the timescale of the problem at hand. On long timescales, we believe that a model based on the same mechanics as disease spread would not provide accurate insight into local distribution of language and would not track native language propagation. This led us to adopt our current model.

Our final algorithm involves iteratively recalculating the local population distributions, $\rho_{x\ell}$, each year by finding the populations of each country sorted by language after applying our model. Our algorithm takes in values for the growth rate r, carrying capacity k, emigration rate e_x, and initial population N for each country being considered in the model. The model then outputs expected populations and expected language distributions for each country. In addition, global native speaker populations are tracked and recorded throughout the simulation.

5 - Results

After implementing our model, we did discover some changes within the top ten languages spoken, as well as a somewhat uniform distribution of languages spoken in every country. This altered our view about the best countries to open a new branch as well as possible changes in the future.

5-1 Top 10 Language Results

Our Top 10 results were a combination of our projected increase in the number of native speakers, the emigrating population, as well as the immigrating population. We individually investigated multiple widely developed countries in order to confirm the interaction between these three values.

The most basic output produced is a ranked list of the languages spoken in all countries across the world. Figure 1 features the current ranks of languages in 2015, while Figure 2 features the projected ranks of the top 10 languages in 2065. Although there are relatively few changes, Bengali and French do switch places between 2015 and 2065 The numbers in Figures 1 and 2 are a combination of L1 and L2 (See Appendix A).

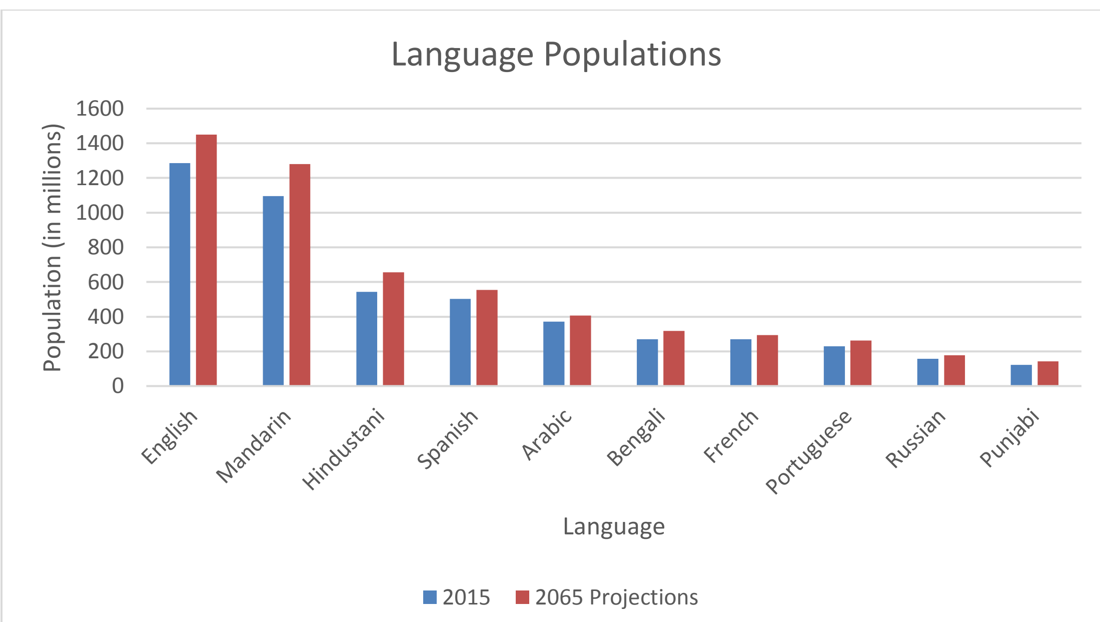


Figure 1

5-2 Geographic Distribution

Currently (as of 2015), the majority of the countries around the world are relatively monolingual, or at least have a few majority languages. As we approach 2065, however, we see a constant pattern of more languages being spoken in more countries. More specifically, the high deviation of current geographic distributions eventually evens out to a smaller population speaking that language in a given country. In Figure 2, data provided by the United Nations suggests that Mandarin Chinese is mostly spoken in China.

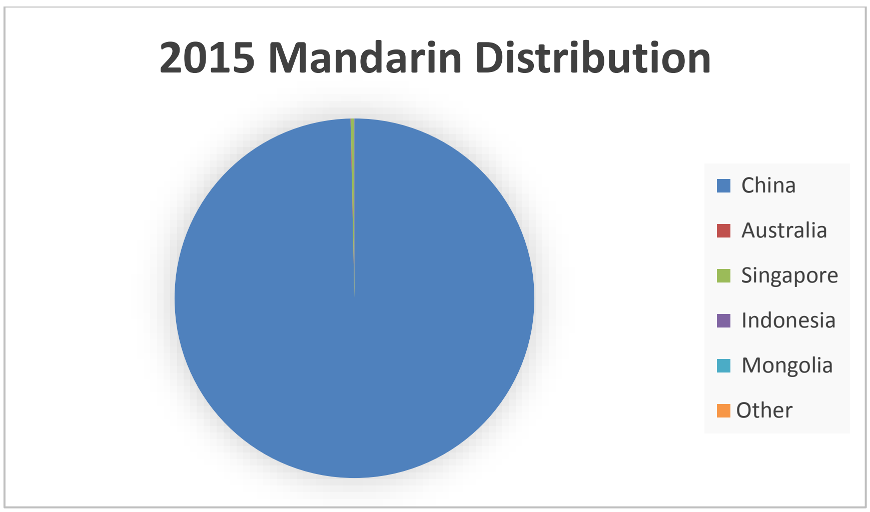


Figure 2

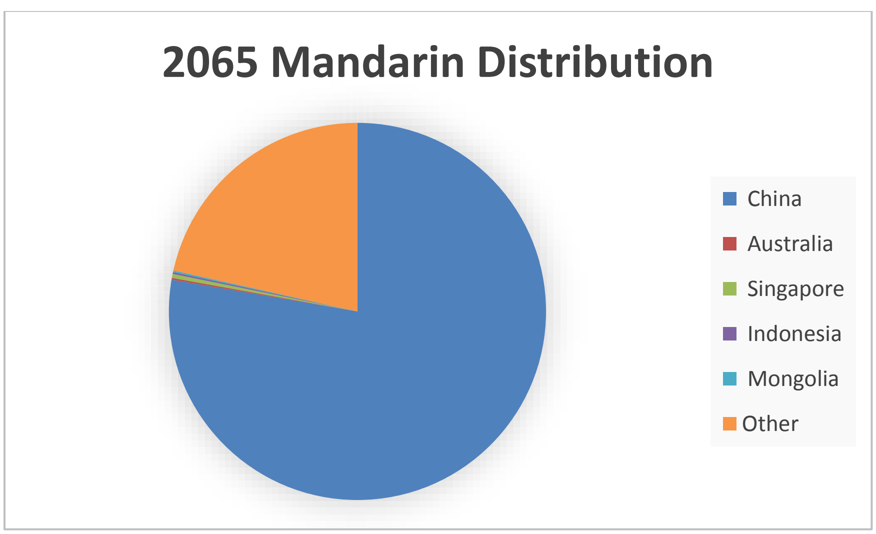


Figure 3

Figure 3 however, uses our model to show that although the population speaking Mandarin Chinese remains primarily in China, there is significant increase of other countries being represented in 2065. Since almost all other countries follow suit in this pattern, we conclude that over time, a language's geographic identity spreads to other countries through migration.

5-3 Office Locations

Initially we sorted the countries by populations and percentage of population projected to speak English. After, we noted the top two languages, apart from English and Mandarin Chinese, in each of the countries from the previous step, and distributed 6 locations to best represent the top ten languages, as well as each continent across the world.

Using this method, we arrived at the following additional six countries: India, Brazil, Russia, France, Japan, and Nigeria (See Appendix B-D). Furthermore, we would construct them in the order in which they were introduced based on the languages they offer from the top ten previously discussed. Figure 4 features a map of the current locations as well as our suggested additions.

Each dot represents addition location

Figure 4

We feel however, that 6 locations are not necessary and would advise 5 locations instead. The Nigeria branch does not add a unique language compared to the other locations and was only suggested to provide a presence in Africa. If the company suddenly finds themselves spending excess resources providing their services in Africa, we do suggest an additional branch in Nigeria (See Appendix E).

If the company experiences extreme growth worldwide and needs to add an additional branch, we suggest installing it in a country with an existing branch, such as Los Angeles in The United States, or another location in China.

6 - Sensitivity

When analyzing the sensitivity of our model, we opted to look at the contribution of each of the modules, population growth, migration, and internal or external influences; and determined their relative effect on the conclusions that were drawn. To evaluate these contributions, we compared the top ten projected languages and the relative populations of the countries tracked.

6.1 Population Growth

Population growth is an important factor to consider when projecting spoken language distributions, at least in the near future (up to 100 years). While it is widely believed that human population growth is limited despite technological advances, populations have yet to react significantly to that limit. In our model, population growth has two major parameters, the intrinsic growth rate r and the carrying capacity k. We will discuss these separately.

The growth rate for a human population after removing all external factors is r. Because of the nature of this parameter, it is impossible to measure and difficult to estimate on any meaningful scale. It should also be noted that the r value of any two human populations would be the same, e.g. $r_{India} = r_{Germany}$. In our testing, we found that $5 \times 10^{-4} \leq r \leq 1.5 \times 10^{-1}$ yielded the same conclusions as our initial growth rate, $r = 1.39 \times 10^{-2}$.

Further, the carrying capacity k of a country is the limiting population size that country can support. Again, this is a parameter that is difficult to set and experts estimating the carrying capacity do not agree. In our model, we assume that the global carrying capacity is proportionally distributed among countries based on current population size. To test the sensitivity of this parameter, we adjusted the global carrying capacity and observed the effects on the top ten languages spoken in the world as well as population distribution among countries. We found that expected global carry capacity values between 1 billion and 500 billion people resulted in similar conclusions for top ten languages and the proportion of global population in each country. As is expected, however, changes in k greatly affected the projected total populations of countries and language speakers.

6.2 Migration

When predicting future geographic language distribution, considering migration rates is a key factor in predicting how people speaking different languages move around the world. In our model we assumed that the number of emigrants from a country in a given year is proportional to the population of that country, represented by e_x. We also assumed that emigrants had equal probability to relocate to each other country. These assumptions were made due to lack of data. To test the sensitivity of our migration model, we varied the value of e and found that the top ten languages did not change for $1 \times 10^{-4} \leq e \leq 2 \times 10^{-2}$ compared to our estimation of $e = 0.005$. However, the distribution of population across the world changes drastically, as higher values of e cause a more uniform distribution of population across countries.

6.3 Global Influences

When considering global influences on language such as government policy, business and international commerce, technological presence, and social pressures; we handled them as a cumulative weight for each language. This weight was applied to the proportion of the population in a given country that did not already know that language. It was also aimed to simulate the desire of the people to learn new secondary languages. Roughly speaking, these values represent the percentage of the population who did not speak a given language will learn that language in the coming year. However, it is difficult to estimate these values. For the purposes of our initial investigation, we assumed that the combined weight of these factors averaged to zero.

We then ran a series of trials with various configurations of randomized influence rates. With reasonable constraints on the bounds of these weights ($\pm 0.05\%$) we found that the order of the top ten languages list did not change and that the total population of languages spoken was not greatly altered. However, at higher ranges of random values ($\pm 5\%$) we found that large inconsistencies in the top ten list and distributions of languages approached 100%.

6.4 Sorting Location Selection

Instead of sorting our candidates by population first, followed by English speaking percentage, we reversed this to analyze the results. Using a sort of English first led to the top 3 candidates with French as their second highest language. This suggests that English and French are highly correlated, and thus give us less accurate results. Since French was highly represented in the top 10 candidates, other important languages in the top 10 were underrepresented.

7 - Conclusion

7.1 Future Additions

Looking forward, we would like model the parameters ourselves, even though we received them from reputable sources such as the United Nations. A key example is the estimate we used for carrying capacity of the Earth. Although we used a common estimate of 11.4 Billion people, there is certainly an opportunity to use Rational Function Regression on collected data about projected growth rates worldwide to estimate another value for carrying capacity, which in turn we would add to the model.

Not only would we further pursue our parameters, but we would also collect more precise data. The limiting factor throughout our investigation was the lack of consistent and reliable data since the problem at hand is at a global level. Many records for languages spoken in small countries are estimations made 5 or even 10 years ago and may plague the accuracy of a long-term model.

Since there was a lack of reliable data, we were unable to accurately come up with unique influence factors on L2 speakers. This field is specifically difficult to measure since the strength of an influence for each individual in each type of situation is subjective. Although we attempted to use several products, specifically games, that have been translated into each other language, this remained relatively ineffective due to the sheer scope that our makeshift data lacked. Thus, we were forced to add an assumption related to the interaction of influence worldwide.

Additionally, we would eliminate as many of our assumptions as possible, since some of them were simply introduced to cope with a time constraint. Assumptions such as ignoring a country's language population under 100,000 people will likely not change the top 10 spoken languages but will greatly impact the final number of speakers worldwide.

7.2 Closing Remarks

Using our model, we determined that although the top 10 languages being spoken in 2065 were the same as those in 2015, their order was slightly different, as to be expected. Also, we discovered that as time continues, geographic language identities tend to fade away, and instead each country holds a more diverse conglomerate of languages.

Furthermore, we introduced and backed our candidates for addition branches and even proposed saving the company money by eliminating an unnecessary branch. Additional and more specific numerical results may be found in the appendix.

MEMORANDUM

TO: Chief Operating Officer

FROM: Team #93797

DATE: 12 February 2018

SUBJECT: New Location Proposal

Greetings from Team #93797!

Although you asked us to provide you with 6 countries, we believe a transition with 5 addition branches would be as effective as 6 branches but reduce the overall cost. Each branch selected has a high representation of English, as well as total population. The following locations are those we selected, in order of audience size and language representation.

1-India 2-Brazil 3-Russia 4-France 5-Japan

We believe India and Brazil are essential in representing Southern America and more diverse Asian audience. After that, we project Russia and France to both contribute to a unique audience in their respective region. Japan offers a key exporting location and another language audience yet to be introduced with the current location spread.

If you find in the future that a presence in Africa is essential in expanding the international presence of this company, we do suggest a location within Nigeria, as its population and language diversity is nearly unrivaled in Africa.

In addition, we further support India and Japan specifically because of International Monetary Fund's projected GDP up to 2022. Indian and Japan join the United States and China in the top 4 highest GDPs worldwide.

If, however, you find that more branches are needed but presence in Africa is not, we suggest adding a second branch in either The United States or China. Both nations have such clear advantages over others in fields such as language diversity, economic stability, population, education background, etc.

Although we understand the time constraints involved in business decisions, if you were to give us more time to investigate this situation we could accurately investigate specific candidate cities to consider. However, if this is not an option we completely understand.

Either way we thank you for this opportunity and hope you find our investigation helpful in your upcoming decisions.

Sincerely,

Team #93797

Works Cited

"Chinese, Wu." Ethnologue, www.ethnologue.com/language/wuu.

"Chinese, Yue." Ethnologue, www.ethnologue.com/language/yue.

"Geographical Distribution of French Speakers." Wikipedia, Wikimedia Foundation, 10 Feb. 2018, en.wikipedia.org/wiki/Geographical_distribution_of_French_speakers.

"German, Standard." Ethnologue, www.ethnologue.com/language/deu.

"Hispanophone." Wikipedia, Wikimedia Foundation, 10 Feb. 2018, en.wikipedia.org/wiki/Hispanophone.

"Italian Language." Wikipedia, Wikimedia Foundation, 7 Feb. 2018, en.wikipedia.org/wiki/Italian_language#Geographic_distribution.

"Japanese." Ethnologue, www.ethnologue.com/language/jpn.

"Javanese People." Wikipedia, Wikimedia Foundation, 8 Feb. 2018, en.wikipedia.org/wiki/Javanese_people.

"Korean." Ethnologue, www.ethnologue.com/language/kor.

"List of Countries by English-Speaking Population." Wikipedia, Wikimedia Foundation, 10 Feb. 2018, en.wikipedia.org/wiki/List_of_countries_by_English-speaking_population.

"List of Countries by Past and Projected GDP (PPP)." Wikipedia, Wikimedia Foundation, 5 Feb. 2018, en.wikipedia.org/wiki/List_of_countries_by_past_and_projected_GDP_(PPP)#IMF_estimates_between_2020_and_2022.

"List of Languages by Number of Native Speakers in India." Wikipedia, Wikimedia Foundation, 6 Feb. 2018, en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India.

"Logistic Equation." From Wolfram MathWorld, mathworld.wolfram.com/LogisticEquation.html.

"Marathi." Ethnologue, www.ethnologue.com/language/mar.

"Punjabi Language." Wikipedia, Wikimedia Foundation, 8 Feb. 2018, en.wikipedia.org/wiki/Punjabi_language.

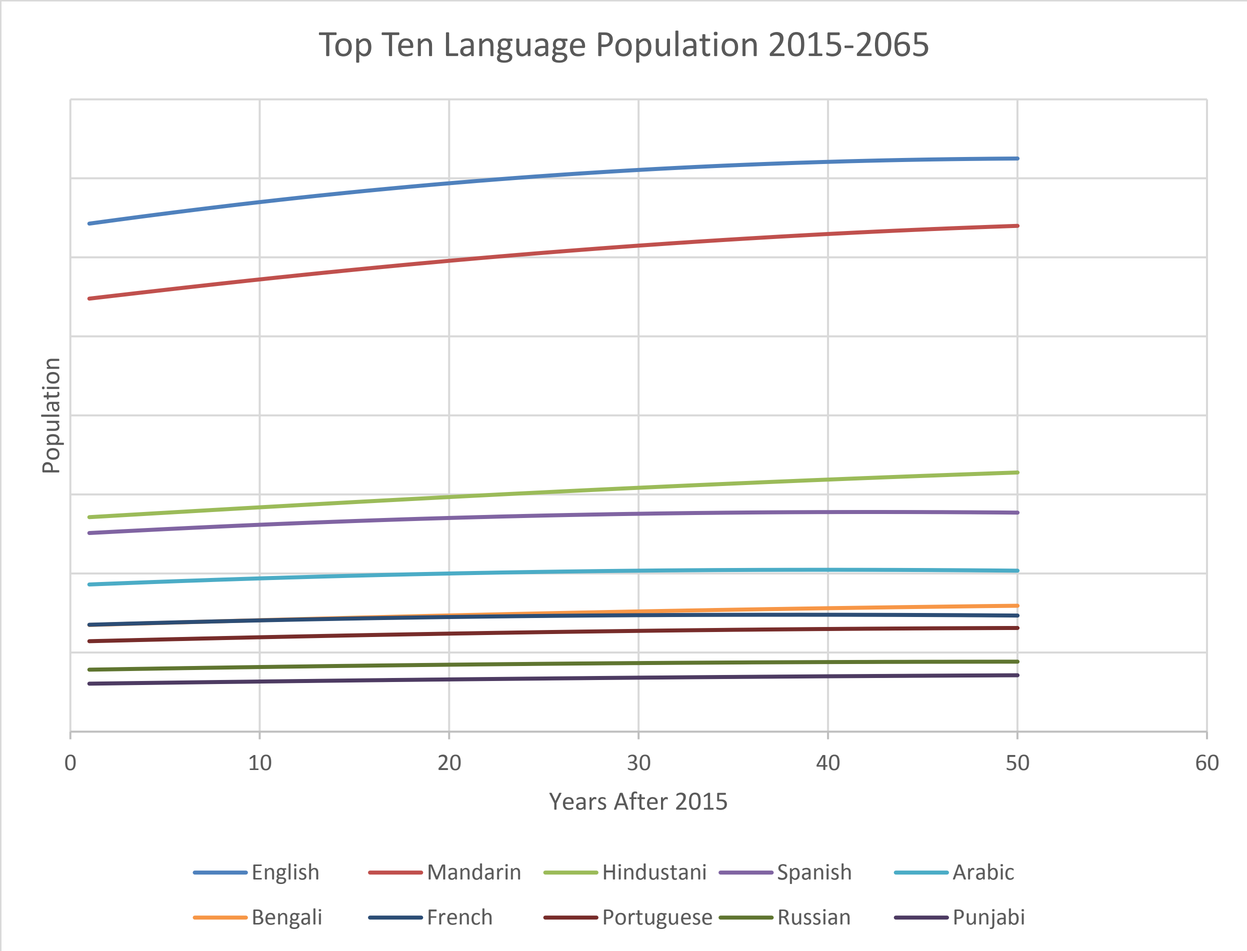
"States of India by Telugu Speakers." Wikipedia, Wikimedia Foundation, 18 Dec. 2016, en.wikipedia.org/wiki/States_of_India_by_Telugu_speakers.

"Swahili." Ethnologue, www.ethnologue.com/language/swh.

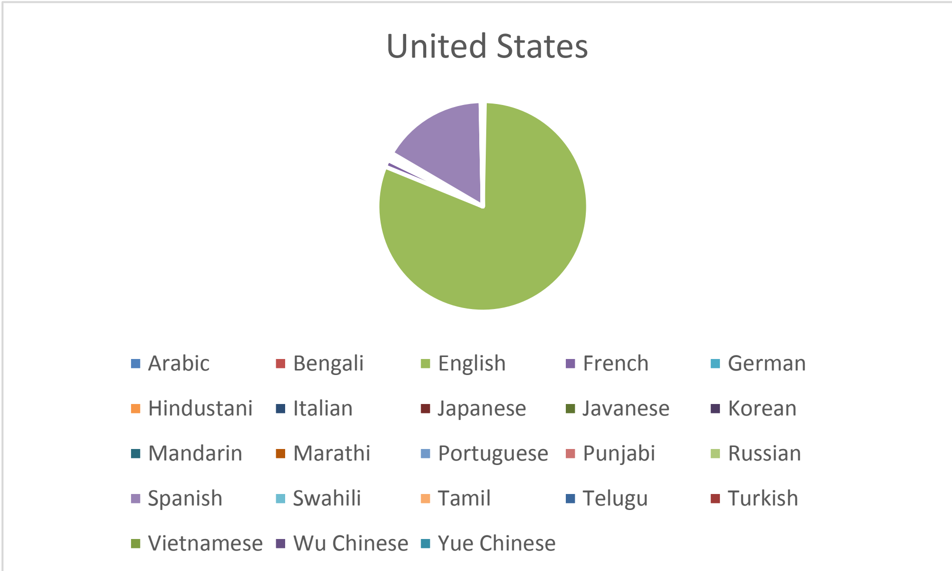
"Tamil Population by Nation." Wikipedia, Wikimedia Foundation, 10 Feb. 2018, en.wikipedia.org/wiki/Tamil_population_by_nation.

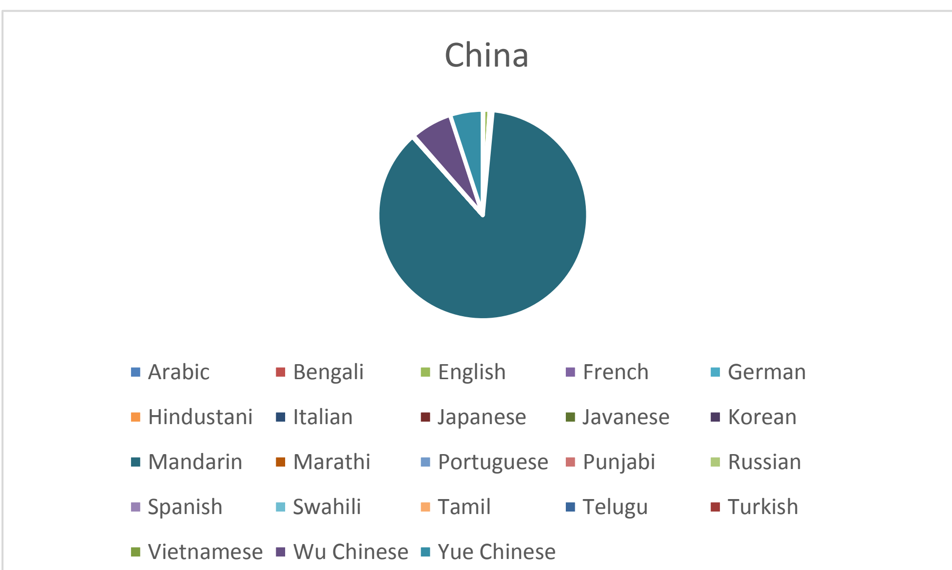
"Turkish." Ethnologue, www.ethnologue.com/language/tur.

Appendix A

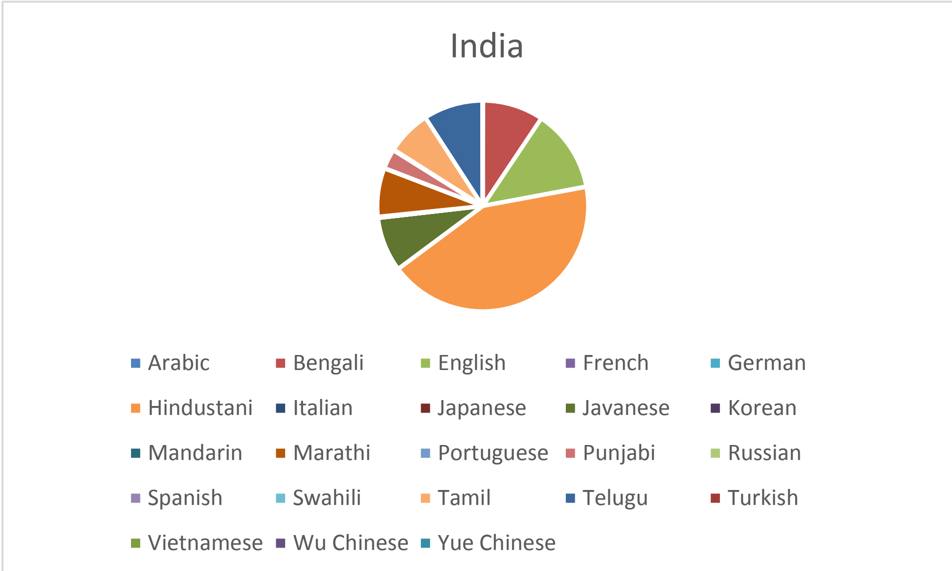

Top Ten Language Population 2015-2065
Population
0
10
20
30
40
50
60
Years After 2015
English
Mandarin
Hindustani
Spanish
Arabic
Bengali
French
Portuguese
Russian
Punjabi

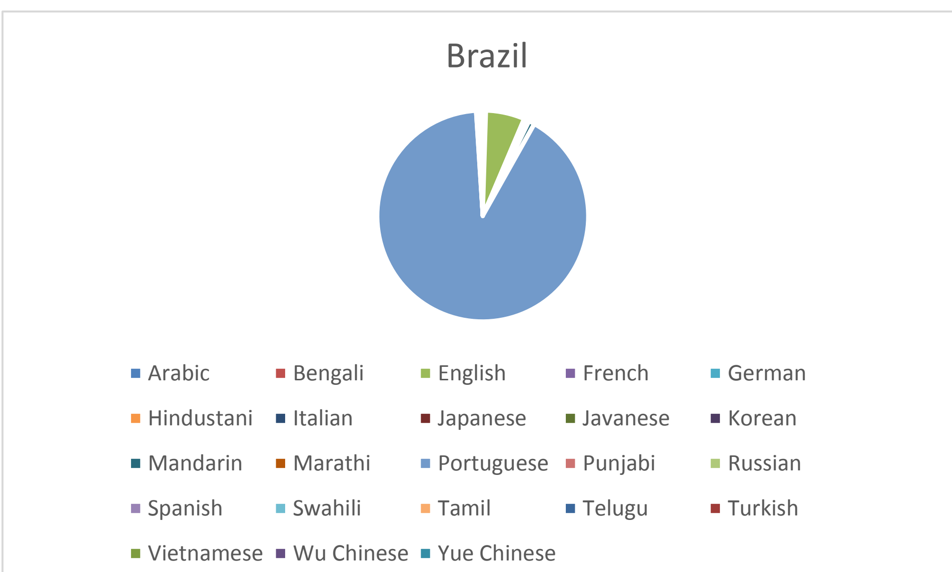
Appendix B


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

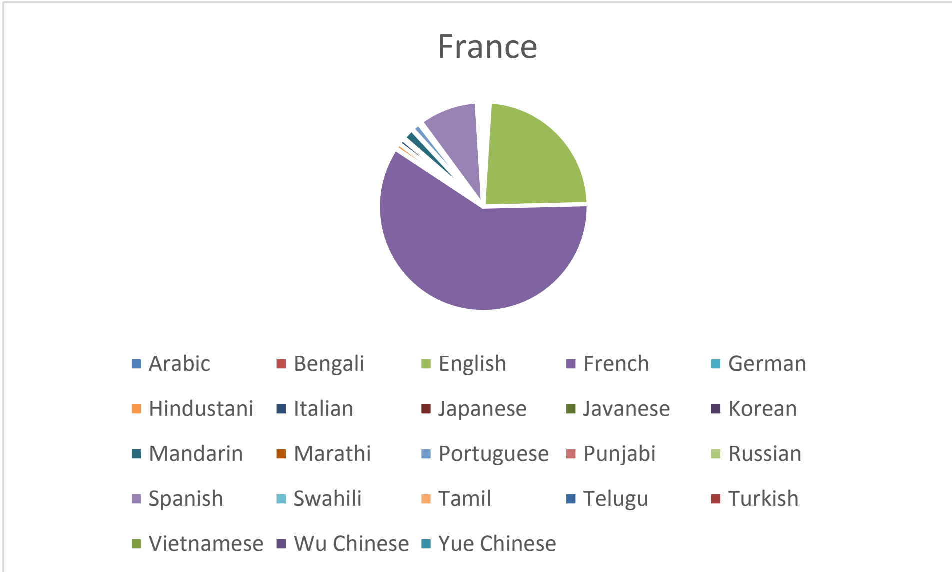

China
Arabic
Bengali
English
French
German
Hindustani
Italian
Japanese
Javanese
Korean
Mandarin
Marathi
Portuguese
Punjabi
Russian
Spanish
Swahili
Tamil
Telugu
Turkish
Vietnamese
Wu Chinese
Yue Chinese

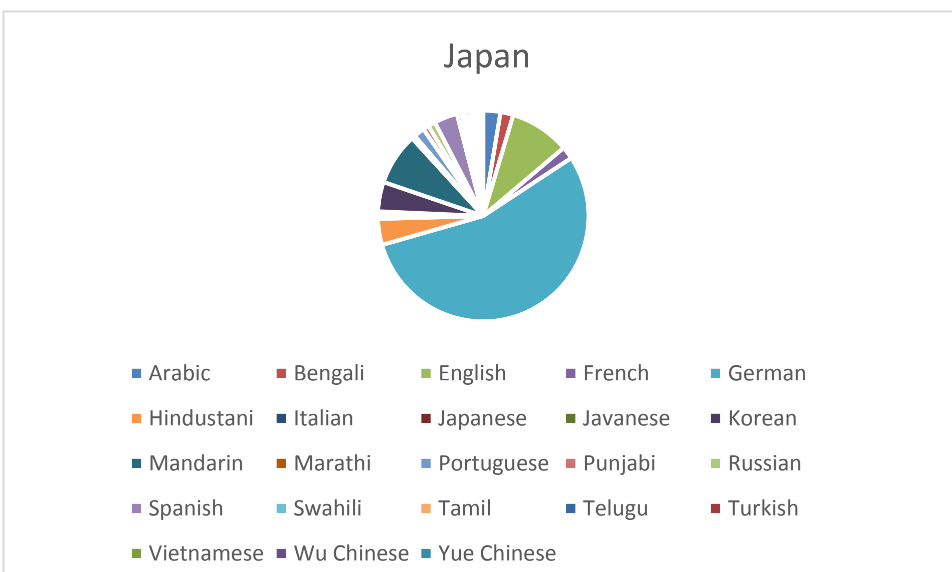
Appendix C


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

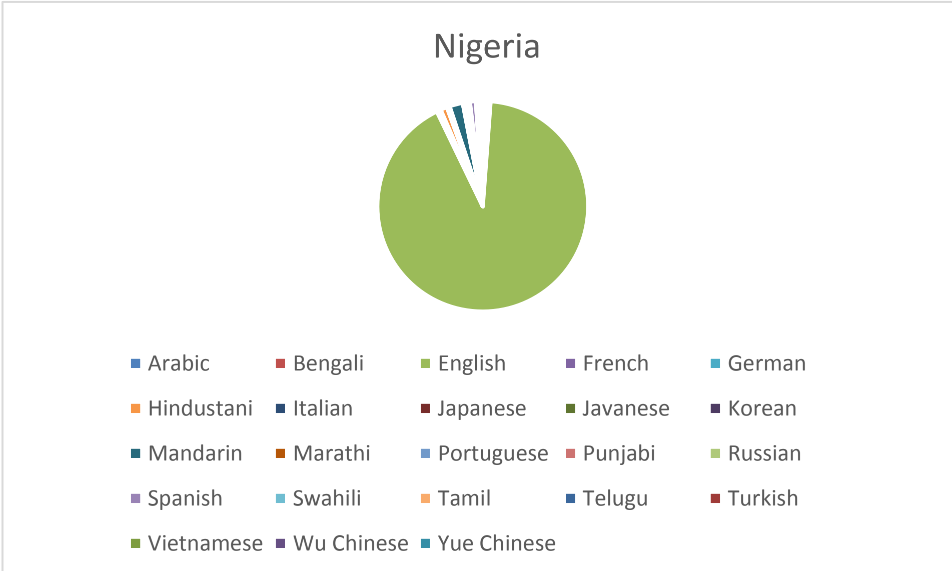

Brazil
Arabic
Bengali
English
French
German
Hindustani
Italian
Japanese
Javanese
Korean
Mandarin
Marathi
Portuguese
Punjabi
Russian
Spanish
Swahili
Tamil
Telugu
Turkish
Vietnamese
Wu Chinese
Yue Chinese

Appendix D


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


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

Appendix E


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