For office use only		Team Control Number	For office use only	
T1	________________	88549	F1	________________
T2	________________		F2	________________
T3	________________	Problem Chosen	F3	________________
T4	________________	B	F4	________________

2017
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.

One of the most critical aspects of growing an international business is for it to adhere to the local culture and customs. One of the best ways by which this can be accomplished is by hiring those who speak the local language. Our task is to determine the future of languages by considering distributions over time and geographic distributions. By knowing this, we can inform the service company which hired us where their next locations for global expansion might be.

In order to predict how many people in the future will speak a given language, we decided to take a statistical approach to the problem. In order to make this approach feasible, we first selected a representative sample of the world in terms of major languages spoken, which we assumed would be understood by modern speakers of the language after 50 years have passed. First, we had to collect and collate a wide array of data concerning the total number and native number of speakers of languages, the population growth ratios, and the migration rates of approximately 100 countries. With this data, we made linear regression-based projections of numbers of speakers of the Top 15 languages in the world today. We checked the regression model against real-world data and used the error to correct the model.

With this corrected model, we predict that when 2070 comes, Hausa will displace Russian from the Top 10 Languages in the World for total speakers. With regard to native speakers, Hausa and French will displace Russian and Japanese for Top 10 Languages. Geographically, English experience dramatic growth in proportion of English speakers in the world in Nigera and Tanzania, whereas Spanish becomes more concentrated in the USA.

We then developed an index of cities based on results of our model. From this index, we specify six cities that the company might benefit from expanding towards; this list includes Mexico City and Dubai.

Language Use and Business

MCM Contest Question B

Team # 88459

February 13, 2018

Contents

1 Introduction: What is the Language of the Future?

1.1 Overview of Forecasting

Forecasting of any process is always to be taken with a grain of salt, especially in the case of predicting linguistic demographics. As with any forecasting activity, determining change and the causes of change can, at best, give the company who wants the forecast some warning signs to look for in the coming years, as well as how to make policies or actions in anticipation of such trends.

1.2 History of Predicting Languages of the Future

1.2.1 Defining Language

Defining languages is a tricky challenge; much research in studying the evolution of languages over time have considered the rules which underlie the use of a language as interacting agents that compete and go extinct, similar to species undergoing biological evolution.

There is controversy as to whether or not languages can be mixed, which implies uncertainty as to whether or not one ought to consider languages as closed structures that only borrow elements when they mutate into different structures. Some think of languages as open to modification while others stick to the languages families constructed.

David Graddol considers how rural languages may die off and new hybrid ones arise in the cities, and that ultimately the traditional notions of language will all but evaporate. [4]. He also considers two models that predict the future of the distribution of language speakers in the world, over time: the Hooke model and engco model. However, these are inaccessible to the public. Furthermore, the usage of scenario planning to make these claims is generally questionable due to their limited use to experts.

1.2.2 Predicting Languages in the Future

Other statisticians who have had a hand at this business, David Graddol and consultants at a French bank, Natixis, have made predictions of the composition of the Top 10 Languages of the world. David Graddol, by futurological technique of scenario planning, thinks children who speak Arabic will outpace those who speak English. [5] French is expected to surge, per Natixia's report.[3]

1.3 Interpretation of Problem

Previous efforts in the mathematical modeling of sociolinguistics have included using predator-prey like models [8, 7, 1] (treating languages as fixed, idealized elements), ala Abrams and Strogatz, as well as using agent-based models to simulate dynamics of agents that represent different grammatical and syntactical rules. However the scale of the problem we consider makes such deterministic models fruitless to use, and the small data available for languages per country make parameter estimation per country somewhat unreliable.

To restate the problem: there is a forecast we must make, which requires the use of statistics and domain knowledge for guess-work. In particular, our predictions, according to Groddol, will be good up until 2100, by when major languages like English and Spanish will undergo major transformations in the structure and globality; i.e. many dialects will come up.

Unlike past attempts at predicting the outcomes of competing languages, our team underwent a statistical approach by amassing (surprisingly!) sparse data sets and collating them to account for the competition between more than two languages, which previous models considered. We then regress number of language speakers over time with respect to some number of factors, and use the results from these regressions to determine how the top ten languages make our recommendations to cities the company that hired us ought to build at.

2 Setup for the Models

2.1 Definitions

- Native speaker of a language: someone who was raised in a country who, by de jure or by de facto, speaks the language.

- Non-native speaker of a language: someone who wasn't raised on a language but learns it to the same degree a native speaker does.

- Language composition: the different languages which members of a particular country speak.

2.2 Overall Assumptions

- Want to say: no new languages will arise. More specifically, languages are open to modification and mixing while still retaining the label.

- Considerations of regional factors like economic prosperity and social stratification change in (a predictable manner?)

- Deformation of a language over time or hybridization won't change dramatically in the span of fifty years.

- No new languages (outlined by Groddel) will arise to displace the ones which currently exist. i.e. structure of current languages doesn't change too much over 50 years from today

- We only approach national languages; this would be in the interest of businesses.

- In the short term, the number of people who speak a particular language is not affected by changes in population.

- Each person has equal influence on who speaks which language.

- Rate of proficiency is the same, regardless of country.

2.3 Variables

- t is time of ten years (i.e. 2020, 2030). Our $t = 0$ is 2020.

- $N(L_i^t)$: number of people who speak language i at some time; indexed in the list in this Section. "base number"

- $N(L_{ik}^t)$: number of people who speak language i in country k at some time t. "base number of country k"

- $P(L_{ik})$: proportion of people in a country k who speak language i.

- M_k^t: population of country k at time t
- $N(L_{ik}) = P(L_{ik}) \cdot M_k^t$
- $emig_k$ emmigration percentage out of some country k, indexed by list this Section.
- $immig_k$ immigration percentage for some country k.
- C_{pro} the proportion of people who become proficient in a language.

3 Model

3.1 Regression on Representative Countries

We assume that the US's change in languages learned represents what other countries of a similar HDI experience. (Intended effect is to take a sample as representative of a larger population, in particular, language competition in other countries.)

So we can look at how changes in the languages specified in the following dataset and generalize to countries with similar HDI.

Collected data from sources on US college language skills. Source: World Factbook, Wikipedia.

We hypothesize that a more complex model will take the following factors into account:

- Population growth per country
- emigration and immigration rates
- language composition percentages

By following per country values of people who speak languages without explicitly considering other factors of different kinds, like economic and technological, we can talk about changes in language speaking over time strictly due to changing numbers of populations per country, as well as how globalization affects language speaking distributions, per physical migration values.

Additionally, the main idea for this model is that, for the growth of speakers per language is country-dependent. That is, the growth rates of different countries would be critical to determining future numbers. Availability of data was also critical.

$$N(L_i^{t+1}) = \sum_{k=1}^{n} \big((N(L_{ik}) \cdot P_k) + (emmig_k \cdot P(L_{ik})) + P(L_{ik}) \cdot C_{pro} \cdot immig_k\big)$$

The update in a language's speakers is every ten years. (represented by t.) The left-hand-side indicates linear regression of data on migration and population growth over n different countries.

3.1.1 Assumptions

- People emigrating will bring their languages to the target country.
- The language composition of emigrants is the same as that of their home country. i.e. there are no additional selecting factors on why emigrants depart from the country that affect language composition.
- Some proportion of immigrants flee to a target country long enough to be fluent in the major languages of a country.
- The growth of a country's population occurs independently of language spoken.
- We need only to consider countries at extremes of emigration and of migration.
- For languages that are almost exclusive to one country, the growth rate of the population of people speaking that language is representative. (i.e. Russian)
- Countries of origin and countries of destination have similar language compositions, though not necessarily the same proportions for the compositions.
- We assume that the top 20 countries with highest immigration rate take in people from top 20 countries with highest emigration rate.
- We assume that the same model would be used if we scaled up the number of countries whose data we use.

3.2 Model Design

Due to the forecasting nature of the problem, we decided to approach it via statistical inference. However, we were plagued by data that was sparse and hard to mine. As a first attempt, we chose to implement multiple linear regression models for each country, which were done in R.

We decided to look at the Top 15 Languages in the world in order to see how the composition of the Top 10 Languages List would change as time progresses.

Sources of Data: [2]

3.3 Model Assessment

Due to the global scope of our model, we had to focus on a few factors. Once we finished making the model, we checked used the population growth rates of the model to scale up data on the language composition of the world from Ethnologue 20th edition, which tabulates $N(L_i^{2017})$, in order to determine errors in the prediction of $N(L_i^{2020})$. For most languages the estimate was within a few million speakers, which is a small relative error, but other languages had much larger relative errors. We realized that this was due to the assumption we made of ignoring countries where $N(L_{ik})$ was small.

In order to take additional countries into account without collecting additional data, we acknowledge that there's some p_i, which denotes the proportion of countries below a certain population, which tells us how much of the actual number of language speakers we're not accounting for. $N(L_i^t) = p_i \cdot N^*(L_i^t)$, where $N^*(L_i^t)$ is the full population of language speakers at time t. Our earlier assumption about growth rates of countries tells us that growth of language speakers in the smaller countries would be similarly calculated for smaller countries, had we processed that data.

4 Results

4.1 Part I. Language Demographics Over Time

4.1.1 Part I.A. + I.B. Distribution of Language Speakers over Time

Per Table 1, in terms of total speakers, by the time 2070 comes around, we can see that English will dominate as the language of the future, with Arabic, French, and Hausa also experiencing large increases in number of

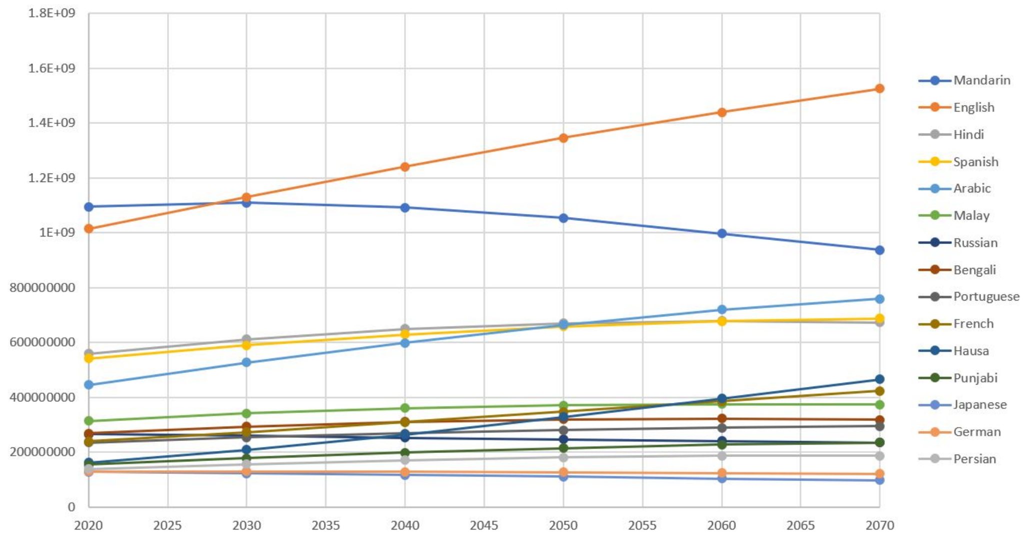

Figure 1:

speakers and increases in ranking. Meanwhile, Russian falls off of the Top 10 list by a considerable amount due to their slowly declining population.

An interesting observation about the rankings is that increasing the number of language speakers doesn't guarantee an increase in ranking, as one can see from the orange entries of Table 1.

Per Table 2, the composition of the Top 10 Languages in the world will change, for native speakers. The languages that drop off the Top 10 list are Russian and Japanese, which are replaced by Hausa and French. Due to population number and migration trends, some of the remaining members of the Top 10 List in 2017 reappear in 2070 in different spots.

Table 1:

Ranking of Total Speakers in 2017 to 2070			
Language	**2017**	**2070**	**Change in Placement**
Mandarin	1 (1.1)	2 (0.95)	-1 (-0.15)
English	2 (0.983)	1 (1.6)	+1 (+0.617)
Hindi	3 (0.544)	5 (0.69)	-2 (+0.146)
Spanish	4 (0.527)	4 (0.78)	0 (+0.253)
Arabic	5 (0.422)	3 (0.8)	+2 (+0.378)
Malay	6 (0.281)	8 (0.39)	-2 (+0.109)
Russian	7 (0.267)	12 (0.233)	-5 (-0.034)
Bengali	8 (0.261)	9 (0.333)	-1 (+0.072)
Portuguese	9 (0.229)	10 (0.295)	-1 (+0.066)
French	10 (0.229)	7 (0.428)	+3 (+0.199)
Hausa	11 (0.15)	6 (0.467)	+5 (+0.317)
Punjabi	12 (0.148)	11 (0.247)	+1 (+0.103)
Japanese	13 (0.129)	15 (0.096)	-2 (-0.03)
German	14 (0.129)	14 (0.122)	0 (-0.007)
Persian	15 (0.121)	13 (0.191)	2 (+0.07)

4.1.2 Part I.C. Geographic Distributions Over Time

Geographically, English experience dramatic growth in proportion of English speakers in the world in Nigera and Tanzania, whereas Spanish becomes more concentrated in the USA.

Figures 2 and 3 show the geographic distribution as a change in concentration of the total number of speakers of English; the bar on the bottom represents percentage of speakers of the language in the title (for the aforemntioned figures, English) of the plot by color. Additional distributions for the other fourteen languages we considered are in the Appendix.

Table 2:

Native Speakers Ranking (Number in billions)

Language	2017	2070	Change in Placement
Mandarin	1 (0.897)	1 (0.77)	0 (-0.127)
Spanish	2 (0.436)	3 (0.569)	-1 (+0.133)
English	3 (0.371)	2 (0.58)	+1 (+0.209)
Hindi	4 (0.329)	5 (0.406)	-5 (+0.077)
Arabic	5 (0.290)	4 (0.521)	+1 (+0.231)
Bengali	6 (0.242)	6 (0..294)	0 (+0.052)
Portuguese	7 (0.218)	7 (0.281)	0 (+0.063)
Russian	8 (0.153)	11 (0.134)	-3 (-0.019)
Punjabi	9 (0.148)	9 (0.234)	0 (+0.086)
Japanese	10 (0.128)	13 (0.096)	-3 (-0.032)
Hausa	11 (0.085)	8 (0.264)	+3 (+0.179)
Malay	15 (0.077)	12 (0.102)	+3 (+0.025)
French	17 (0.076)	10 (0.14)	+7 (+0.064)
German	18 (0.076)	<15 (0.071)	<0 (-0.006)
Persian	25 (0.060)	14 (0.093)	+11 (+0.033)

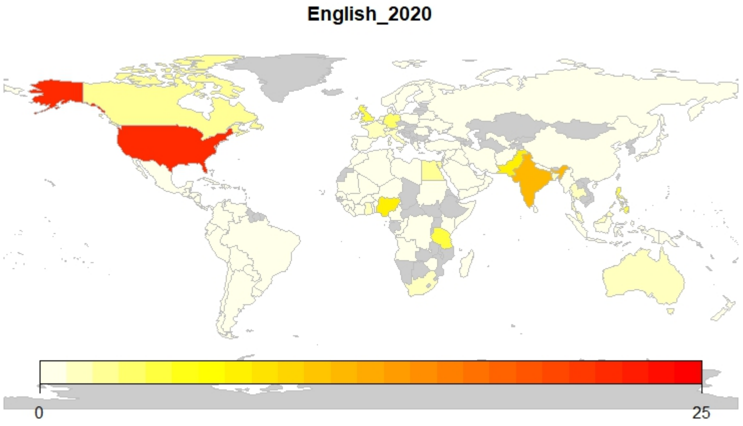

Figure 2:

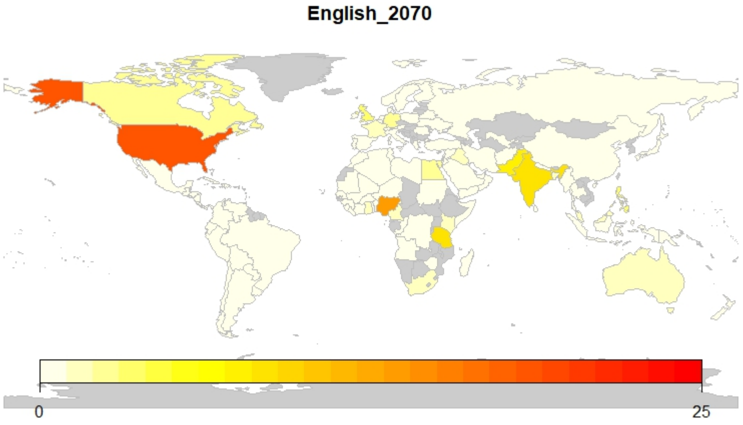

Figure 3:

4.2 Part II. Recommendations for Office Locations

4.2.1 Part II.A. Six New Offices

We made an index of which cities as follows:

$$I = \frac{50}{100}LR + \frac{35}{100}GLR + \frac{10}{100}LCR + \frac{5}{100}EPR,$$

where

- LR is the Language Rating, (determined by model)
  - Based on our results of language distributions, English will grow to be quite ubiquitous, so searching for people who are fluent in a non-English language will weigh the most.
- GLR is the GLobalization Rating, (determined by Global Integration Rating)
  - The idea for this is that the more connected a city is with the world, the more desirable it is for business.
- LCR is Living Cost Rating, (determined by Living Cost)
  - The cost of running a business will affect profit.
- EPR is the English Proficiency Rating (determined by EF English Proficiency Index).
  - Due to the demand of the service company that hired us, we consider English proficiency as a salient factor for the index, but our previous results show that it won't be hard to find people proficient in English for certain countries.

Based on the index calculated for 2017 and for 2070, the cities we recommend are Mumbai, Mexico City, Beijing, Dubai, Hong Kong, and Singapore. From the present into the future, these six cities remain the top ones into which a company seeking to expand its influence ought to go. (See Table 3). While we have yet to test this index out fully, the justifications we provide for each rating used in the index indicate that it approximately matches the desires of the service company. Due to how any Hausa-speaking city ranks too low for our index, despite the growth in the number of speakers of that language, we do not recommend African cities.

Table 3:

City	Official Language	2017 Rank	2070 Rank	Potential
Beijing	Mandarin	1	1	0
HongKong	Mandarin, English	2	2	0
Singapore	Mandarin, English, Malay	3	3	0
Dubai	Arabic	4	4	0
Mumbai	Hindustani	5	5	0
Mexico City	Spanish	6	6	0
Madrid	Spanish	7	7	0
Kuala Lumpur	Malay	8	9	-1
Moscow	Russian	9	12	-3
Paris	French	10	8	2
Sao Paulo	Portuguese	11	11	0
Brussels	French, Dutch	12	10	2
Tokyo	Japanese	13	13	0
Frankfurt	German	14	14	0

4.2.2 Part II.B. The Potential of Opening Fewer Offices

As has been previously discussed in the Introduction, a combination of interactions between people, specifically young ones [5], and technologies like the Internet [4] and telecommunication will ultimately drive English into a variety of dialect-like tongues throughout the world. In addition, urbanization is expected to drive the formation of hybrid languages, which would mean that entirely new cultures, and hence entirely new markets, would arise. If we could keep track of the demographics of these languages with transformed rules over time, rather than just associating languages with a nation, we might be able to specify fewer than six locations through a similar regression analysis done in this report.

5 Part III. A Memo to the Chief Operating Officer

Results and recommendations regarding the trends of global languages and location options

We investigated the trend of the global languages by using historical growth data and immigration effects, then build a model that can predict the growth of global languages together with their demographic locations. For the next 50 years, starting from 2020, there are some unexpected changes in global languages. English is having a consistent growth over time and surpassed Mandarin as the most spoken language in the world. In the other hand, Mandarin has an increasing trend up to 2030, then starting to decrease until 2070 as the population of China decreases, but Mandarin still will manage to stay being the second most spoken language in the world, mainly due to the enormous number of population in China. Another interesting change is made by Arabic which will jump to the third place surpassing Spanish (4th) and Hindi (5th) that have less growth rate. French also will make a significant jump from 10th to 7th, but the one that gives the biggest surprise is Hausa, who started the 2017 survey at 11th place and ends up at the 6th place surpassing 5 languages including French. Some possible explanations of these phenomenon are the migrations directions and different population growth speeds in a certain area in the world.

We also investigated the change in demographic distribution of each languages from 2020 to 2070 to see how the heatmaps of language distributions change in 50 years. For most languages, the results are not surprising, because they are spoken in some specific countries of which not many people will migrate out. However, some noticeable changes will happen to English, Spanish, and French. The distribution of English in the United States will decrease and shift towards countries in the Africa, which is unexpected, and this might show the direction in which the global economy might grow. For Spanish, the heatmap moves from Mexico to the United States, possibly explained by historical migration from Mexico to the United States, which had established the foundation for Spanish speakers in the United States. Lastly, the direction of growth of French speaker leads to Africa as well, especially in Congo. This is mainly due to the high population growth expected in Africa in the next 50 years.

With regard to recommendations for new office locations, we have taken into consideration various factors, including language, global integration level, English proficiency level in the countries, and cost. After creating an index based on importance rankings of these factors, we have made a list of the top 20 cities for possible new office locations, and the top 6 choices are the followings: Mumbai, Beijing, Hong Kong, Singapore, Dubai, Mexico City.

For the long term, we will need the data from demographics so that we could keep track of the demographics of these languages over time. Also, if we can get access to updated data of factors mentioned above, we may have a more accurate prediction of top megacities in the future, as well as number of offices needed.

6 Future Work

As with any time series forecasting, having more clean data to increase sample sizes would make these forecasts less uncertain. Had we more time, we would have considered additional factors like climate, urbanization, and so on. The employment of additional methods like non-linear models, grey-box methods, exponential smoothing, ARIMA, and others [6] to forecast the future of language at the scale of the globe will have to be done.

7 Conclusion

Figuring out the future of the languages spoken by humanity is no small task. Country-specific data regarding potential factors for the change in number of speakers of a particular language over time must be gathered and assessed for as many countries as possible in order to get a reasonably accurate picture of change in the distribution of speakers of a particular language over time.

Despite the sparseness of such data, we developed a linear regression model that, with our assumptions and adjustments, takes a representative sample of countries in the world, population growth, migration, and language composition to make predictions of what the language of the future might be. The figures above indicate that the composition of the Top 10 Languages in the world, for both native and total speakers, will include the Hausa language. Additionally, our geographic distributions indicate that English and Spanish will spread into new countries more quickly than other languages will.

The upshot of the model presented in this paper is how it takes into account far more data than prior studies have, which only looked at English or a few other languages at most.

References

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

[2] CIA. The World Factbook: Languages.

[3] P.-E. Gobry. Want To Know The Language Of The Future? The Data Suggests It Could Be...French, Mar 2014.

[4] D. Graddol. The Future of English? *London: The British Council*, 7:9–23, 1997.

[5] D. Graddol. The Future of Language. *Science*, 303(5662):1329–1331, 2004.

[6] R. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2nd edition, May 2017.

[7] M. Patriarca, X. Castell, J. R. Uriarte, V. M. Eguluz, and M. S. Miguel. Modeling two-language competition dynamics. 2012.

[8] T. Raducha and T. Gubiec. Predicting language diversity with complex network. *CoRR*, abs/1704.08359, 2017.

Appendix

Part I.C. Figures

All of the numbers at the right-end of the colour-bar are percentages; the percentages of speakers of a particular language in each country sum to 1.

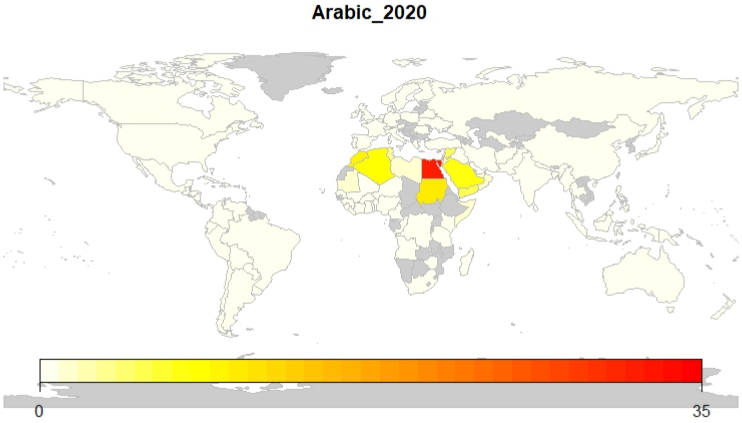

Figure 4:

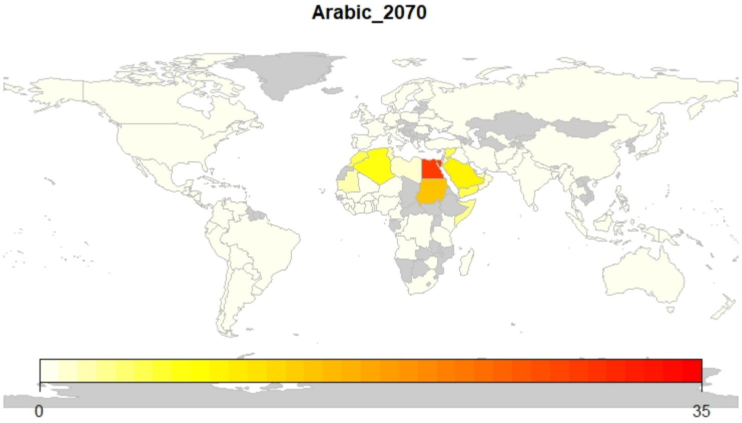

Figure 5:

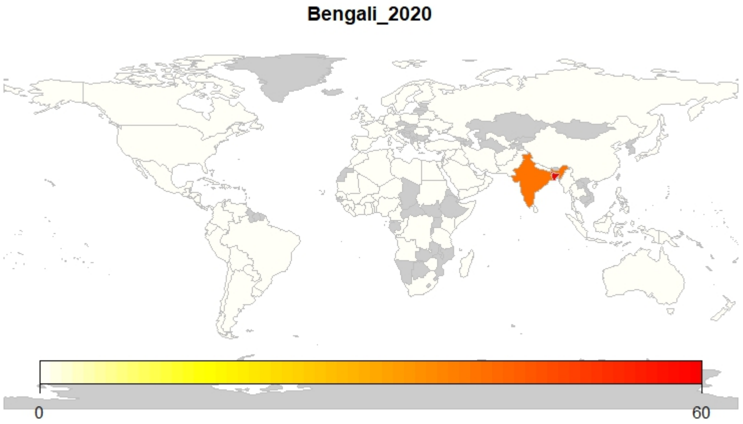

Figure 6:

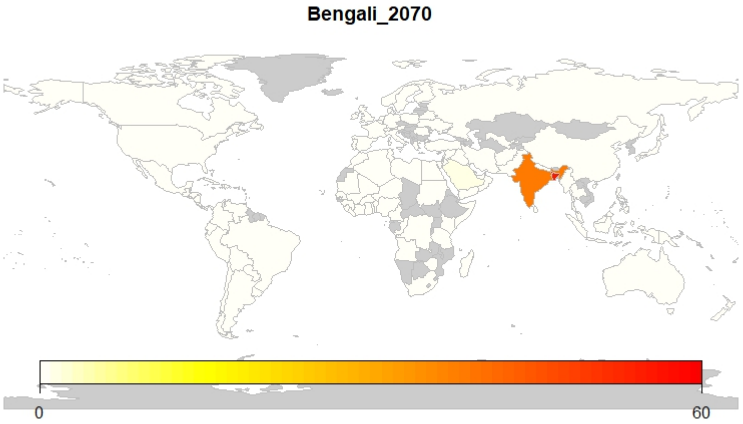

Figure 7:

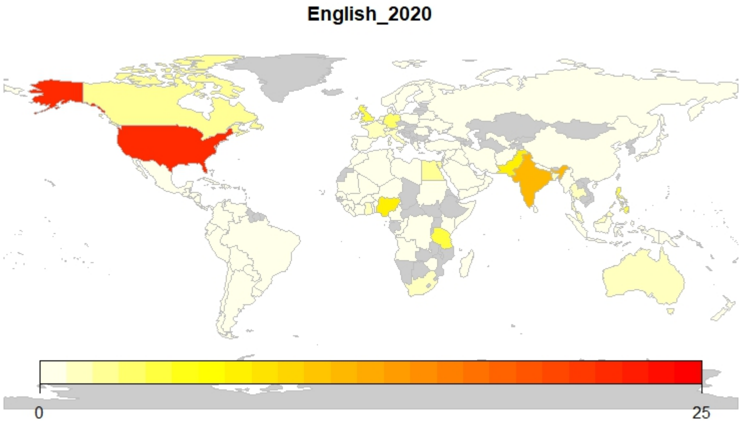

Figure 8:

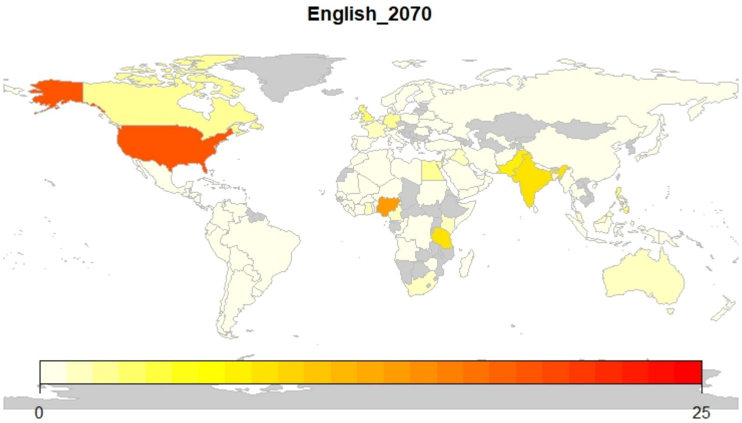

Figure 9:

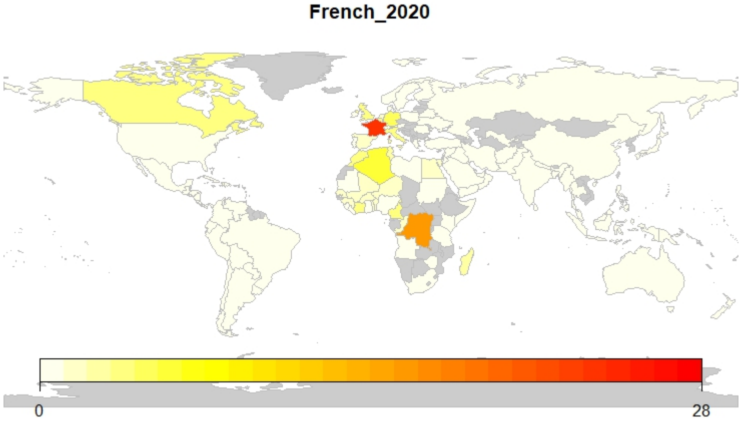

Figure 10:

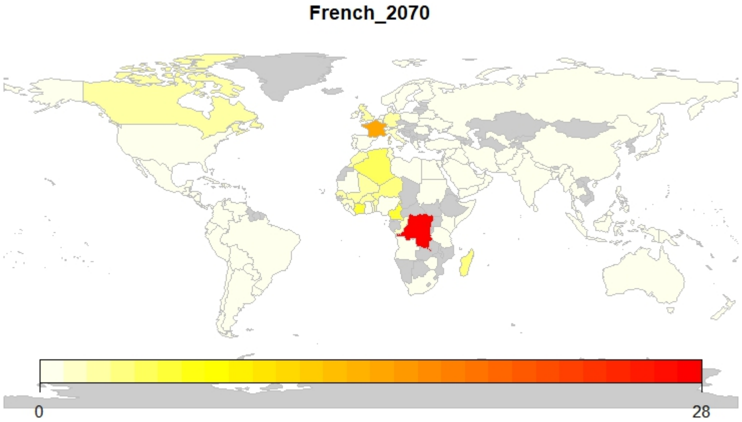

Figure 11:

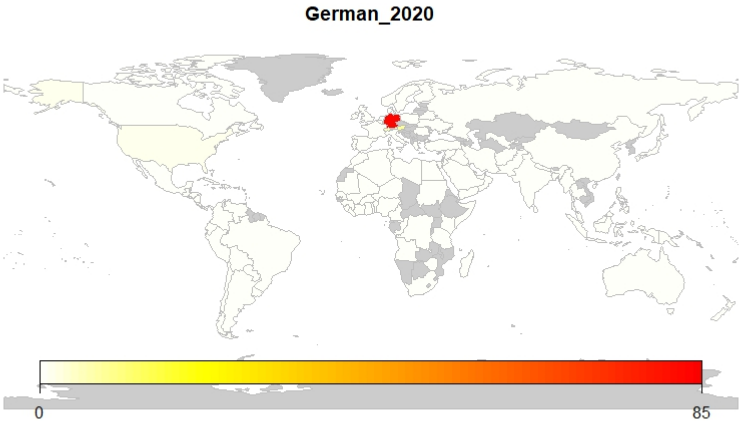

Figure 12:

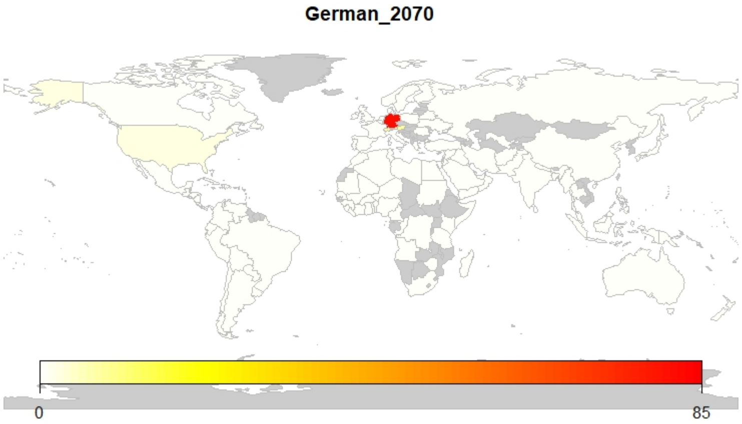

Figure 13:

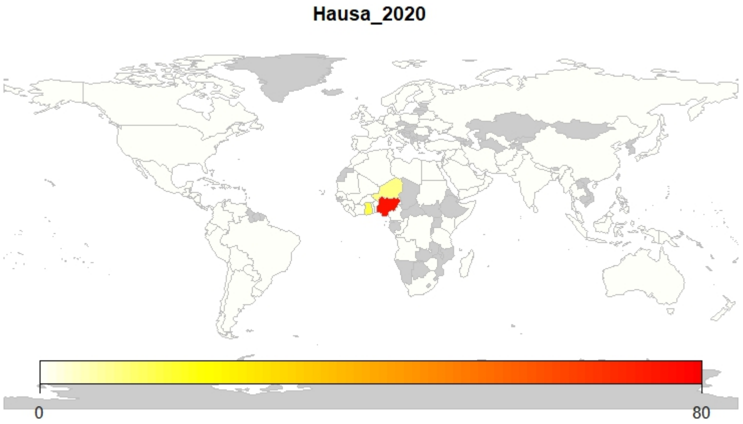

Figure 14:

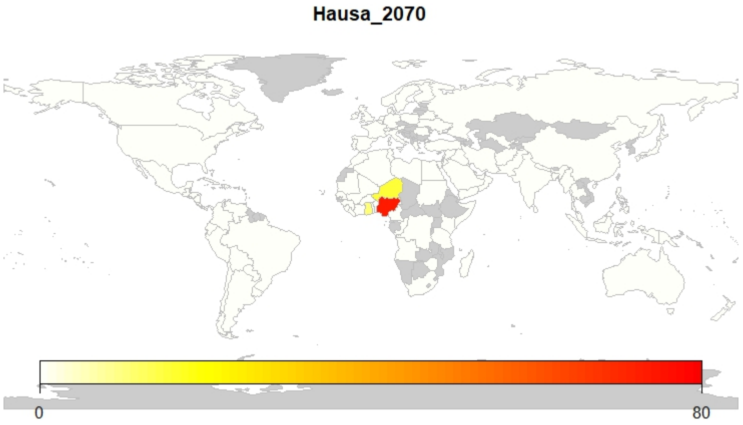

Figure 15:

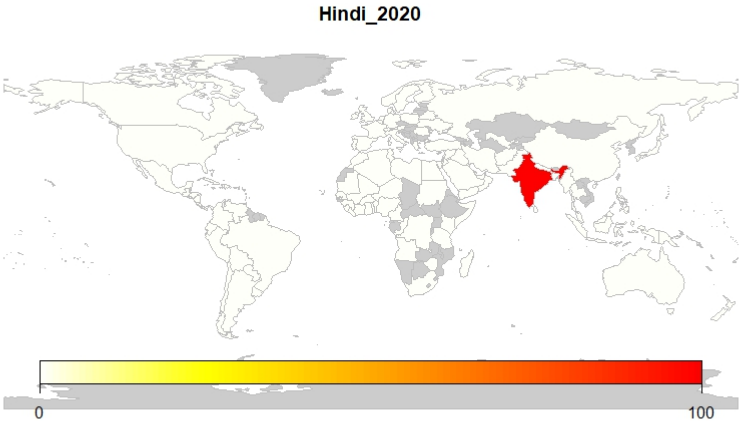

Figure 16:

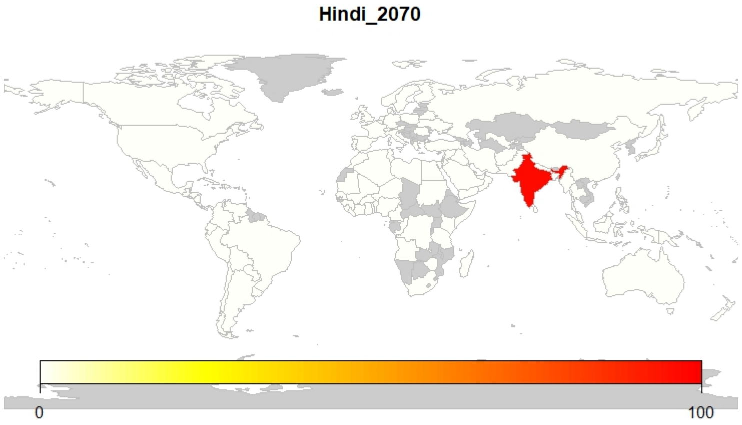

Figure 17:

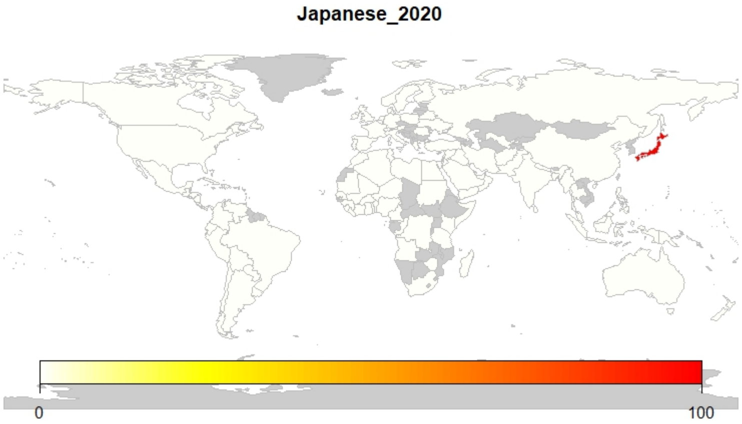

Figure 18:

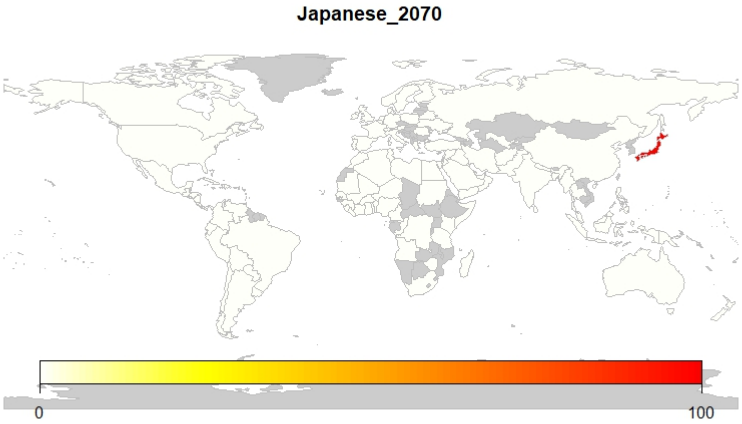

Figure 19:

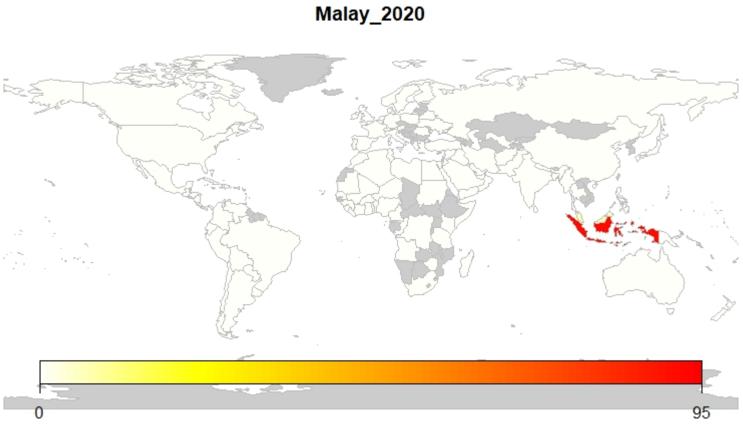

Figure 20:

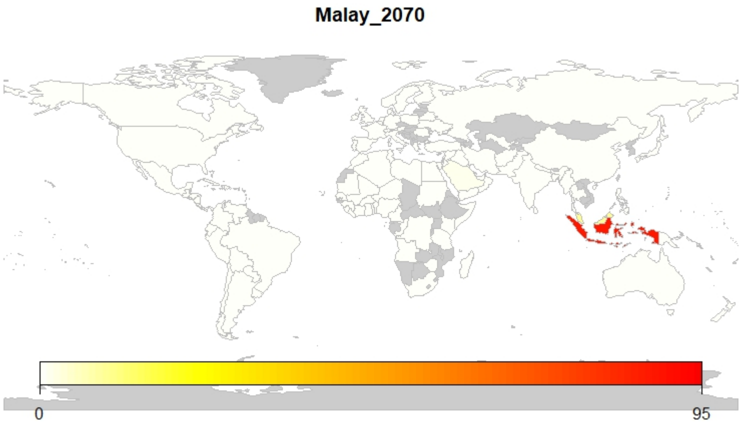

Figure 21:

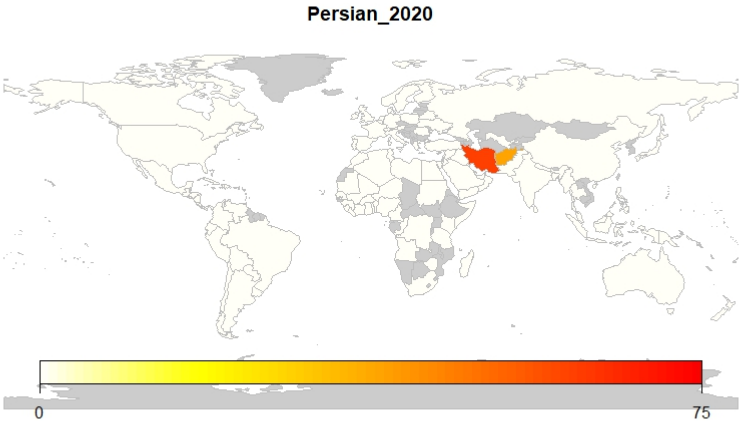

Figure 22:

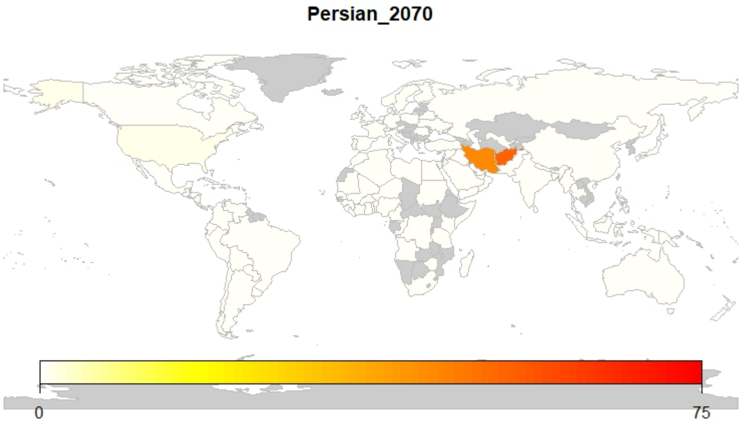

Figure 23:

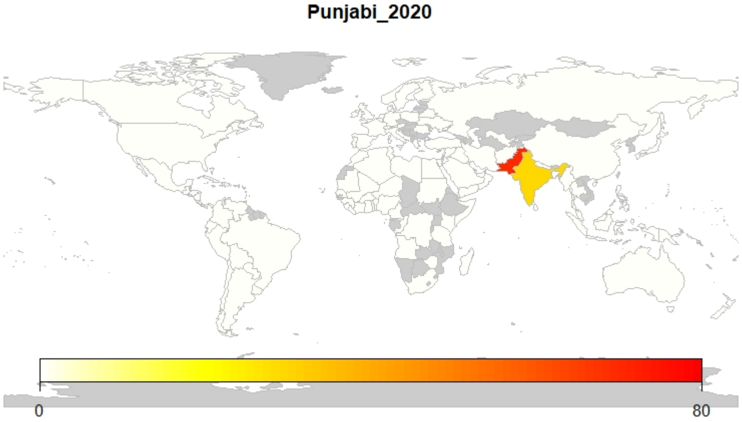

Figure 24:

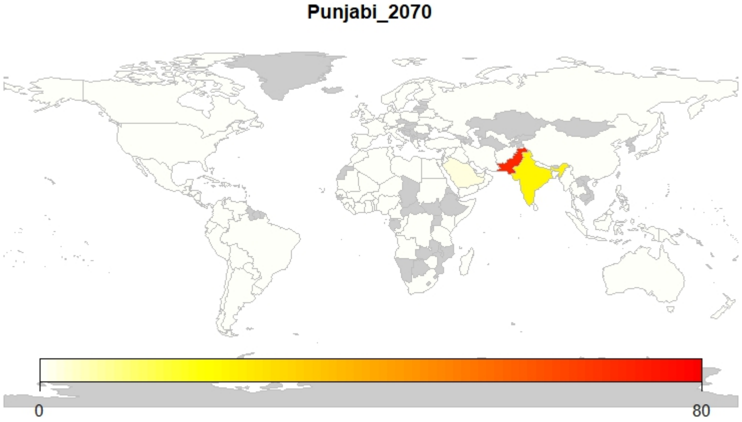

Figure 25:

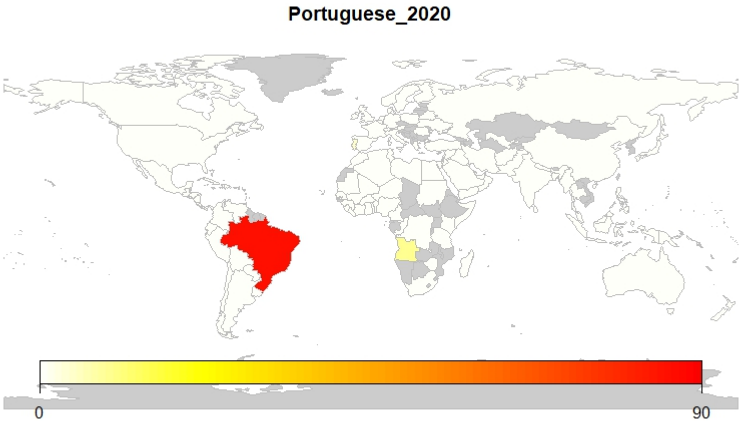

Figure 26:

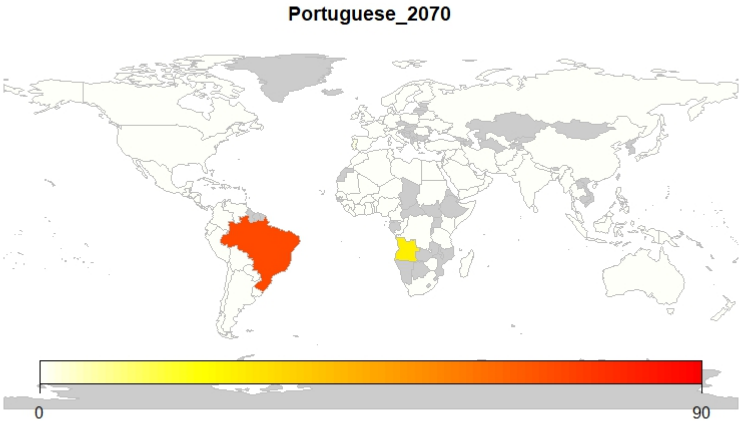

Figure 27:

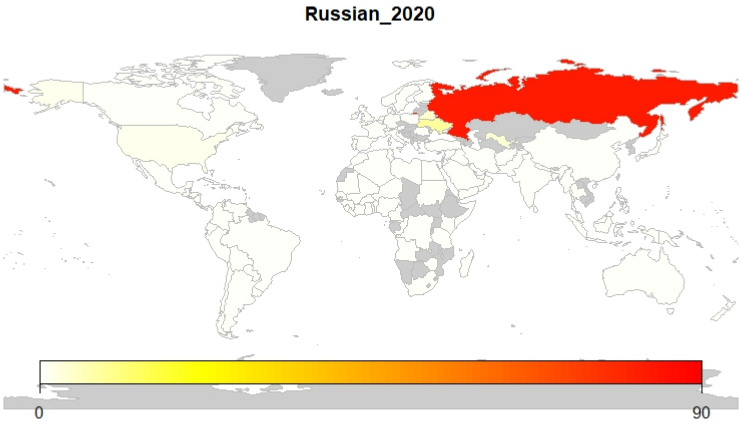

Figure 28:

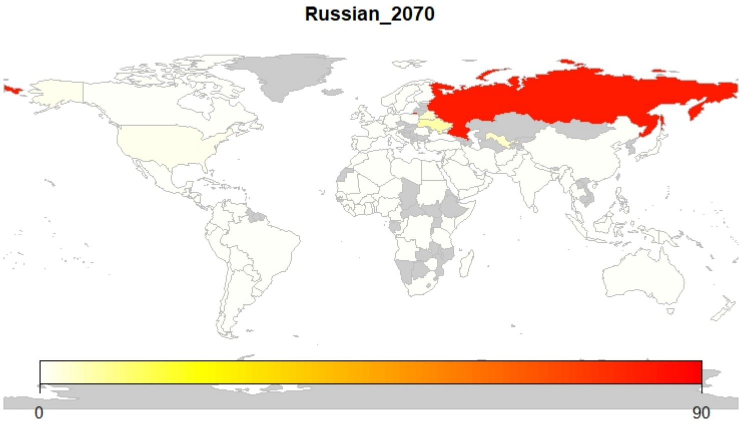

Figure 29:

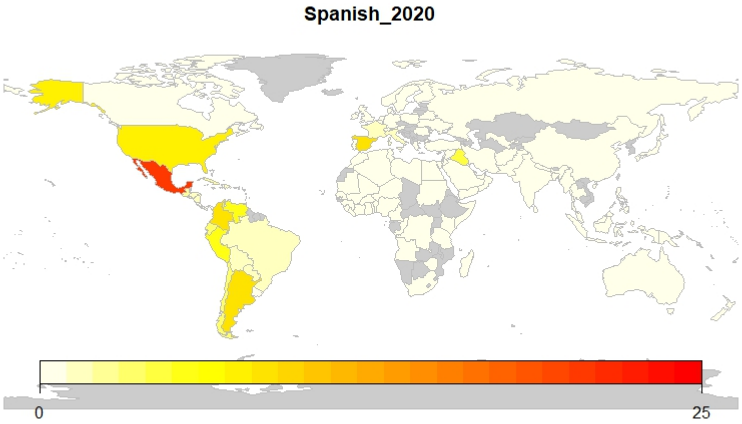

Figure 30:

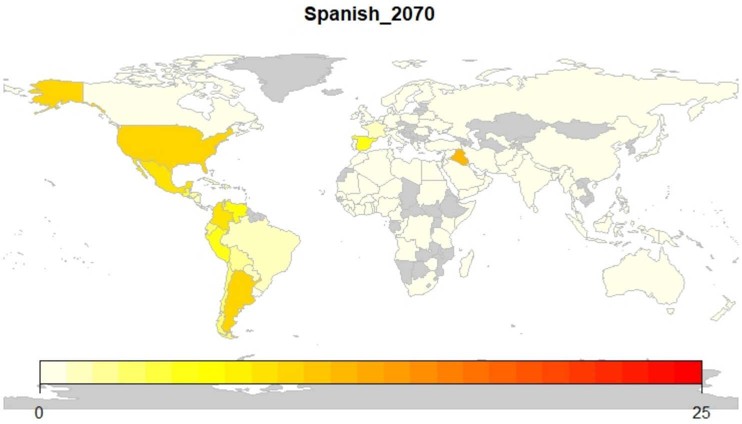

Figure 31: