ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 159

International Conference on Mathematics, Modelling, Simulation and Algorithms (MMSA 2018)

Simulating the Quantitative Variation of Language Speakers under Globalization

Zhen Song, Jiaqian Chen and Xiao Han*

School of Economics, Jinan University, Guangzhou, 510000, China

*Corresponding author

***Abstract*—This paper analyzes the quantity changes of speakers among the major languages in the world. We apply trend extrapolation to forecast native speakers. Meanwhile, we apply Markov chain to simulate the process of migration and learn dynamic changes of non-native speakers. We conclude that Mandarin Chinese ranks first and 63% migrants as language carriers choose the U.S. and thus English maintains its status. We finally list top 10 languages in terms of numbers of native speakers and total speakers.**

Keywords—Markov chain; trend extrapolation; migration; zglobalization

I. INTRODUCTION

Language, as a tool for communication within and among communities, is interrelated with cultural exchanges, economic status or political situation of nations. The reason is that human activity vastly influence language acquisition and development. Besides, a 'widely-spoken' native language in the destination country can be a pull factor in international migration, the core theme of language evolution and the outcome of economic exchange, cultural integration and technical progress worldwide [1]. In this paper, we attempt to predict the quantitative variation of the native language users and non-native users in 50 years by using trend extrapolation and Markov chain. Based on our findings, we can obtain new numbers and rankings of users of language in 2067 and give some discoveries and conclusions.

II. THEORETICAL BASIS

Based on the fact that language users are composed of native and non-native speakers, we will divide the total number of users into two parts:

Native speakers [sociolinguistic definition]-someone speaks a language as first language or mother tongue

Non-native speakers [pedestrian definition]-someone learns a particular language as a child or adult rather than a baby [2]

Total language users = native speakers [L1] + non-native speakers [L2]

For one thing, number of 'Native speakers' is obviously decided by the number of people in the country where the language is spoken. So 'Native speakers' is primarily related to fertility and mortality of the corresponding country.

For another thing, the evolvement of non-native speakers are complexly related to history, social stratification and personal prospect. As a preliminary research, we consider that migration being the most important human activity which change the number of non-native speakers [3]. According to empirical data, people who is the first generation to a country, his/her non-native language generally is the language of the destination country. Second or above generation immigrants aren't considered in this paper.

III. MODELING AND SIMULATION

There are over 5000 languages in the world. Studying every language will make our work huge. So we select top 10 languages in terms of number of L1 speakers and total number of speakers. To analyze the quantity changes, we separately forecast the number of native speakers and non-native speakers. Hence, we could add up to the total number of users and obtain a precedence table. For the reason that the using precondition of Markov Chain is a closed system, we converge the language mentioned above to build a Language Pool which we think the future top 10 languages will derive from it, considering their current strong numerical advantage. We should narrow the scope further.

A. Define Research Subjects

According to the current number of language users, we can easily select altogether 12 languages as follows:

TABEL I. QUANTITY & RANKING OF SPEAKERS OF CHOSEN LANGUAGE IN 2017 [4]

Language	L1	L1 Rank	Total	Total rank
Mandarin Chinese	897	**1**	1090	**1**
Spanish	436	**2**	527	**4**
English	371	**3**	983	**2**
Hindustani	329	**4**	544	**3**
Arabic	290	**5**	422	**5**
Bengali	242	**6**	261	**8**
Portuguese	218	**7**	229	**9**
Russian	153	**8**	267	**7**
Punjabi	148	**9**	148	12
Japanese	128	**10**	129	13
Malay	77	15	281	**6**
French	76	17	229	**10**

According to the data in Table I, the number of L1 speakers is much larger than non-native speakers (mainly L2 speakers), so our previous assumption that number of L1 is mainly decided by the population of nations where above predominant language spoken by native is reasonable. The 12 languages will be the future TOP 10 languages based on the two forceful facts: first, the number of 13rd language speakers is much fewer than that of 12 countries ahead; second, the summation of 12 languages

users occupy more than 50% of the total population in the world.

However, it is still hard to consider all the countries belonged by the 12 languages. Hence, we regard 80% (Representative Rate) as the dividing line. In terms of L1 speakers, if the total number of the language-using countries' citizens to all users' number of this language is over 80% [5], we think these countries could represent all the language-using countries. And then, we could concentrate our attention to research the selected 17 countries. In this way, the research scope is much more specific and representative. By now, our research objects have changed from languages to countries.

TABEL II. LIST OF 7 LANGUAGES JUST BELONG TO ONE COUNTRY

Language	L1	P	PR	Country
Mandarin Chinese	897	897	**100%**	China, **CN**
Hindustani	436	429	**99%**	India, **IN**
Russian	153	138	**90%**	Russia, **RU**
Punjabi	148	148	**100%**	Pakistan, **PK**
Japanese	128	126	**98%**	Japan, **JP**
Malay	77	77	**100%**	Indonesia, **ID**
French	76	65	**85%**	France, **FR**

P: population; PR: proportion. The 7 languages above belong to 7 countries where L1 speakers take account for almost 100%. Therefore, these countries themselves could represent the language coverage area. When languages exist in several countries. we take IMS%, IMS and MTC into consideration to choose countries and produce the next form by collecting data. [6][7]

TABEL III. LIST OF COUNTRIES BELONG TO OTHER 5 LANGUAGES

Language	L1	Country	Pn
Spanish	436	Mexico**, MX**	120
		Columbia	48
		Spain**, ES**	46
		Argentina**, AR**	43
		Peru	30
		Venezuela	29
		Chile	17
		Ecuador	16
PR	**80%**	Sum	349
English	371	America, USA	234
		Britain**, UK**	65
		Canada**, CA**	21
PR	**86%**	Sum	320
Arabic	290	Egypt**, EG**	82
		Algeria	45
		Iraq	39
		Morocco	38
		Sudan**, SD**	28
PR	**80%**	Sum	233
Bengali	242	Bangladesh**, BD**	158
		India**, IN**	83
PR	**100%**	Sum	241
Portuguese	218	Brazil**, BR**	201
		Angola	9
		Mozambique	7
PR	**100%**	Sum	217

First, we ranked the countries by the number of people who speak the language as native language. Then, we ranked the countries by the number of its native speakers and chose the countries of which summation is more than 80%.

$$PR = \frac{\sum Pn}{L1} \geq 80\%$$

Pn: the population of people who took this language as native language

We chose 28 countries and it is still difficult to analyze all of them. So, we conducted the second selection by three indicators below.

International migrant stock (%): IMS% could partly measure the openness and mobility of the country. If IMS% is small, the mobility is low, as to say, the number of non-native speakers of the country is relative stable. Based on our premise, we eliminated trivial IMS% countries.

International migrant stock (total, million): IMS million is a supplement of IMS% index. For instance, the IMS% of a country is relative high, while its citizens are very few and the international migrant stock is so small that will not influence the world pattern at all. We eliminated these 'small migrant countries'.

Migrant stock change (1995-2015): MTC measures the migrant change in recent 20 years, which will contribute to our prediction. We prefer to choose the countries whose value is greater.

Finally, we choose the following 17 countries to further our research:

America, Argentina, Bangladesh, Brazil, Britain, Canada**, China, Egypt, France, India, Indonesia, Japan, Mexico, Pakistan, Spain, Sudan, Russia**

B. Trend Extrapolation For L1 Number

We calculate the number of native speakers (L1) depends on total population and natural growth rate of each country. To that end, we collect the data about population of these 17 countries from 1960 to 2016 to forecast the quantity in 50 years. If the country accords with Linear Growth Model (L), we will use linear fitting equation to estimate. Otherwise, we will take nature growth rate into account to estimate other Non-Linear Growth Model (NL). In the NL, we calculated by the average natural growth rate to predict the population in 50 years. We define the total population as Pn of each country.

On the other hand, not all people in a country speak the official language as a mother language and they may share more than one language, so we calculate the Language Penetration Rate in our selected countries and add is to formula. We use TABLE I- III and get the LPR (Language Penetration Rate) and the number of L1 speakers in 2067. P_{2017} is the population of the country in 2017. The native speakers in each country are calculated by the second equation. So far, we have computed the native speakers (L1) of each country.

$$LPR = \frac{L1}{P_{2017}}$$

$$L_1^n = LPR \times P_n$$

C. Markov Model For Non-Native Speakers

In this section, we apply Markov chain to build initial state and transition matrix to simulate the situation of the migrant and population change in the 17 countries.

1) Initial State: According to our theory model, the chosen 17 countries are regarded as a closed system. In order to measure the contribution rate of each country, we define ISR as the initial state distribution of 17 countries.

$$(immigrant\ share\ rate)\ ISR = \frac{IMS(international\ migrants\ stock)}{\sum_{i=1}^{17} IMS}$$

TABEL IV. ISR OF 17 COUNTRIES

AR	**BD**	**BR**	**CA**	**CN**	**EG**
0.020	0.014	0.007	0.071	0.009	0.004
FR	**IN**	**ID**	**JP**	**MX**	**PK**
0.071	0.047	0.003	0.021	0.011	0.031
RU	**ES**	**SD**	**UK**	**USA**	
0.105	0.053	0.007	0.079	0.448	

Then, we regard the ratio share of migrants in 2017 as the initial distribution.

$$x(t_i) = (x_{i1,} x_{i2,} \cdots x_{i17})\ \ (countries: 1, 2, ..., 17)$$
$$t_i (t = 1, 2 \cdots n)\ \ (years: 1, 2, ... N)$$

2) First-Step State Transition Matrix: The emigrant stock and immigrant stock from each country to the rest 16 countries are regarded as the inflow and outflow.

We use P to remark first-step state transfer matrix:

$$p = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1\bullet 17} \\ p_{21} & p_{22} & \cdots & p_{2\bullet 17} \\ \vdots & \vdots & \vdots & \vdots \\ p_{17\bullet 1} & p_{17\bullet 2} & \cdots & p_{17\bullet 17} \end{pmatrix}$$

$$i: row \quad j: column \quad p_{ij}: probability\ of\ transition$$

The probability indicates that the system transfers from state i to state j. Then, we build up the following equation and attain the first matrix.

$$P_l + P_o = PIP - \sum_{i=1}^{16} P_i$$

PIP: potential immigrant population of the rest 16 countries

P0: citizens who immigrate to countries apart from the 17 countries

Pl: citizens who have the will to immigrate but still at home

Pi: citizens immigrate to one of the 17 countries

We collect the potential migrant ratio of the 17 countries (statistics of Gallup, 2017). So, we construct the migrants' pool for further prediction. [6][7]

$$(potential\ migrant\ ratio)\ PMR = \frac{(Citizens\ want\ to\ immigrate)\ CI}{Total\ popoulation}$$

$$\begin{array}{c} \\ AR \\ BD \\ \cdots \\ UK \\ USA \end{array} \begin{array}{c} \begin{array}{ccccc} AR & BD & \cdots & UK & USA \end{array} \\ \begin{pmatrix} 5545798 & 0 & \cdots & 11339 & 179639 \\ 0 & 21636559 & \cdots & 39803 & 80995 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 943 & 33470 & \cdots & 7192105 & 714999 \\ 5135 & 46008 & \cdots & 212150 & 44091017 \end{pmatrix} \end{array}$$

We define the sum of P_0 and P_1 to be the diagonal values. As the matrix indicates, for instance, $P_{17\bullet 2}$ shows that 46008 people immigrate from USA to BD and $P_{17\bullet 17}$ means in 2017, there are 44091017 citizens who are willing to immigrate still staying in the USA or immigrating to the other countries out of our research system. Based on the matrix, we could gain the flow probability formula:

$$p_{ij} = \frac{x_{ij}}{\sum_{j=1}^{n} x_{ij}}$$

Then, we could compute the first-step state transition matrix:

$$p = \begin{pmatrix} 9.12\times10^{-1} & 0 & \cdots & 1.87\times10^{-3} & 2.96\times10^{-2} \\ 0 & 9.59\times10^{-1} & \cdots & 1.76\times10^{-3} & 3.59\times10^{-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1.03\times10^{-4} & 3.67\times10^{-3} & \cdots & 7.89\times10^{-1} & 7.84\times10^{-2} \\ 1.14\times10^{-4} & 1.02\times10^{-3} & \cdots & 4.72\times10 & 9.81\times10^{-1} \end{pmatrix}$$

3) Preliminary Result: We use $x(t_i) = (x_{i1,} x_{i2,} \cdots x_{i17})$ to represent the distribution of IMS in 17 countries $t_i (t = 1, 2 \cdots n)$. Next, we may get the distribution of IMS:

$$x(t_i + 1) = (y_{i1,} y_{i2,} \cdots y_{i17})$$

Furthermore, we could exert the prediction formula:

$(P_1(1), P_2(1), P_3(1), \cdots P_{17}(1)) = (P_1, P_2, P_3, \cdots P_{17}) \bullet P$ Based on the initial state distribution of the population mobility in the 17 countries and transition probability matrix, we predict the ISR of them and present the result.

$$x(t_i) = (x_{i1,} x_{i2,} \cdots x_{i17})$$

TABEL V. ISR OF 17 COUNTRIES IN 2067

AR	**BD**	**BR**	**CA**	**CN**	**EG**
0.015	0.016	**0.04**	**0.037**	0.023	0.011
FR	**IN**	**ID**	**JP**	**MX**	**PK**
0.039	**0.056**	0.023	**0.026**	0.001	0.019
RU	**ES**	**SD**	**UK**	**USA**	
0.01	0.026	0.004	0.023	**0.629**	

As the table shows, there are 64.79% of people who want to immigrate in our closed system moving to the USA in 2067. In short, the geographical distribution and quantity of language users changes along with migration at the same time. USA will attract a large number of immigrants. Furthermore, non-English

countries tend to use English as L2 language for its international trade and cultural exchange while English countries choose their L2 languages mainly in light of geopolitics and geography.

D. Calculate The Population

We have obtained the population of each country and then we use the PMR (Potential Migration Rate) of the 17 countries to predict CI (citizens want to immigrate). According to the Markov process, we have acquired the estimated value of ISR (Immigrants Share Rate) in 2067. By virtue of these indicators, we could calculate ISN (migrants):

$$CI = P \times PMR$$

$$ISN = CI \times ISR$$

From another point of view, not all people move to another country require to learn a new language on condition that their mother language is also wildly used in this country, or even they tend to immigrate to countries share the same mother language with their hometown [8]. As a result, we also use LPR (Language Penetration Rate) to amend the number of non-native speakers.

$$L_2^n = ISN \times LPR$$

$$Total\ \ L^n = L_1^n + L_2^n$$

TABEL VI. 17 COUNTRIES COMPUTING RESULTS IN 2067

	TE	**P**	LPR	**L1**	**L2**	**Total**
AR	L	66	98%	64	16	80
BD	L	275	97%	267	16	283
BR	L	344	97%	333	42	375
CA	L	52	57%	30	23	53
CN	NL	1780	65%	1160	16	1170
EG	NL	263	86%	226	10	236
FR	L	84	97%	81	41	122
IN	L	2150	6%	135	4	139
			25%	535	15	550
ID	L	423	29%	125	7	132
JP	NL	120	99%	119	28	147
MX	L	211	94%	199	1	200
PK	NL	544	9%	49	2	51
RU	NL	157	96%	150	10	160
ES	NL	46	98%	46	27	73
SD	NL	126	71%	90	3	93
UK	NL	97	98%	95	25	120
USA	L	453	72%	328	486	814

IV. RESULTS

From the last part, we know both the rank of L1 speakers and total rank in 50 years.

TABEL VII. LANGUAGE RANKING IN 2067

Language	**L1**	**L1 rank**	**Total**	**Total rank**
Mandarin Chinese	1157	**1**	1196	**1**
Hindustani	535	**2**	579	**3**
English	453	**3**	850	**2**
Bengali	402	**4**	434	**4**
Portuguese	333	**5**	359	**5**
Arabic	316	**6**	343	**6**
Spanish	309	**7**	337	**7**
Russian	150	**8**	154	**8**
Malay	125	**9**	143	**10**
Japanese	119	**10**	150	**9**
French	81	11	113	11
Punjabi	49	12	53	12

By making a new language rank table in 2067, we find Punjabi will be replaced by Malay compared to top 10 languages in terms of L1 speakers in 2017. Mandarin Chinese is still in the top, while the order of Hindustani and English will slightly change. In light of our model, India will be the biggest country for population instead of China if it maintains the current growth rate. So its L1 speakers will surpass English. English maintains its status in terms of total number because of its vast L2 speakers. Finally, we can forecast the distribution of language in 50 years [9].

ACKNOWLEDGMENT

We thank Dr. Ye for his help in the field of measurements and simulation analysis. We also thank all teachers who have helped us to develop the fundamental academic competence. Last but not least, we would like to thank all friends and families for their encouragement and support.

REFERENCES

[1] Adams P C. (2017) Neutral Accent: How Language, Labor, and Life Become Global. Duke University Press, 2017(1):131-133.

[2] Adserà A & Pytliková M. (2015) The role of language in shaping international migration. Economic Journal, 125(586):F49–F81

[3] Adserà, A., & Pytliková, M. (2016). Language and migration.

[4] 2018 distribution of language speakers https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[5] Nattavudh P. ,Richard V. B., & Jan D. E. (2017) Top incomes and human well-being: Evidence from the Gallup World Poll.

[6] [6] Migration data portal. (2015)

[7] https://migrationdataportal.org/?i=stock_abs_&t=2017

[8] International migration in Pishu database.

[9] http://www.pishu.com.cn/skwx_ps/sublibrary?ID=10704&SiteID=14

[10] Lipowska D & Lipowski (2017) A. Language competition in a population of migrating agents. 95(5-1), 052308.

[11] Raymer J. (2017) Measuring flows of international migration. Iza World of Labor. university (social science edition), 17(1),81-86.