Modeling language competition

Yanxin Liu

Dec 14, 2009

Abstract

As same as biological species, languages have the life cycle of birth, evolve, and death. There are about 6900 languages currently spoken in the world. Unfortunately, 90% of them are facing extinction in 21st century as a result of language competition. The loss of linguistic diversity means the loss of cultural diversity. The language competition can be modeled as a collective phenomena resulting from the interactions of individual language speakers. Modeling language competition may be helpful in preserving some of the endangered languages. I will review three popular models exploring the question at different level of descriptions: The top-down macroscopic model proposed by Adrams and Strogatz (AS model) is based on the differential rate equation; A modified version of AS model incorporating bilingualism; and the bottom-up microscopic model based on computer simulations focusing the individual language speakers and their interactions. Possible applications of each model are suggested.

There are about 6900 languages being spoken in the world with very uneven distribution of number of speakers today [1]. Mandarin Chinese, Spanish, and English are the dominant ones with more than 1.5 billion speakers. However, many of the current existing languages are endangered. 473 of the languages listed in the Ethnologue are classified as nearly extinct, defined as "only a few elderly speaker are still living". Languages are considered as "endangered" when parents are no longer teaching the language to their children and are not using it actively in everyday matters. Although there is ambiguity in the definitions of "endangered" and "nearly extinct", it is clear that hundreds of languages will not be passed on to the next generation [2-4]. Language is a precious part of the human heritage, whose extinction will significantly reduce the culture diversity of human society. It has been estimated that one language is dying out every other week [3]. This astonishing extinction rate is a result of languages competition, namely languages are competing for speakers. The social and economical pressure will favor one or few languages among all the competing ones; eventually, the favored ones will win and reach equilibrium. Others will disappear forever. It is what is happening, but certainly not what we want to see. Recently, physicists began to invading the area of so-called "quantitative linguistics", and modeling the languages competition [5-6]. The goal is to understand the underlying dynamics of language evolution, as well as contribute to the language/culture-preservation effort.

I. Introduction

The problem of modeling language competition is approached by physicists from different level of descriptions: top-down macroscopic models and bottom-up microscopic models. In the macroscopic models, the language is idealized as one object with certain number of speakers. The models focus on the evolution of the number of speaker in the population in terms of first order differential rate equations, ignoring internal structure of the language such as syntax, grammar and their changes. The initial model proposed by Abrams and Strogatz [7] also assumed that there is no spatial or social structure in the populations, in which all speakers are monolingual. The modified model proposed by Mira and Paredes take into account the possibility of being bilingual, and introduced a parameter describing the similarity of two languages [8-10]. In contract, the microscopic models monitor each individual language speaker and their interactions [11-13]. Most of the results and conclusions are based on computer simulations. I will explain both models and their modified version with emphasis on the macroscopic one in sections II-IV. Possible application of models will be discussed in Section V as future projects to the interested ones including myself.

II. Macroscopic models

The first macroscopic model [7] of language competition is introduced by Abrams and Strogatz in 2003 (AS model). The goal to track down the time evolution of the fraction of minority speaker of a language, and extract a linguistic parameter that can identify a endangered language at an early stage from empirical data. The parameter obtained from the model can be used to evaluate the threat of language extinction, so that appropriate action can be taken to persevere the language.

In the AS model, only two languages (X and Y) are considered, which are competing for speakers. The number of speakers (precisely the percentage of people in a population) speaking each language is denoted as x and y, respectively. A few simplifying assumptions have been made in the model:

1. The population size is constant. Each individual only speak one of the two language, namely monolingual: (x+y=1);
2. The population is highly connected, with a uniform spatial and social structure. The individuals interact with each other at the same rate;
3. The switch from one language (for example X) to its competing partner (Y) is due to the "attractiveness" of the competing language;
4. The attractiveness of a language increase with both its number of speakers and its perceived status, denoted as s, which is parameter can be understood as some kind of social or economic advantage that a particular language offered to its speakers. The s_X and s_Y are the relative status for language X and Y, satisfying the relation $s_X+s_Y=1$ as a description of the competition.

The mathematical model is constructed based on first order differential rate equation. Suppose an individual speaker converts from Y to X with a probability $P_{YX}(x, s_X)$ per unit time, where x is the fraction of the population speaking X, and s_X is measure of X's relative status. Likewise, The converts probability per unit from X to Y can be written as $P_{XY}(y, s_Y)$. The rate at which the fraction of population speaking X changes can be introduced based on first order differential rate equation:

$$\frac{dx}{dt} = yP_{YX}(x,s_X) - xP_{XY}(y,s_Y) \qquad (1)$$

There is no need to write down the equation for y because x and y are not independent, rather related through x+y=1. Now we turn our attention to the explicit express for the probability P. From assumption number 4 stated above, they further proposed the functional form of P given by power-law:

$$P_{YX}(x,s_X) = cs_X x^a \qquad (2)$$

likewise,

$$P_{XY}(y,s_Y) = cs_Y y^a = c(1-s_X)(1-x)^a \qquad (3)$$

where a is a parameter that models how the attractiveness of a language scales with the number of its speaks. It has been found unexpectly, based on fitting of historical data, that a is roughly a constant across cultures, with $a=1.31\pm 0.25$. The formulation of the switch probability P have taken into account several facts:

1. No one will adopt a language that has no speakers: $P_{YX}(0,s_X)=P_{XY}(0,s_Y)=0$;
2. No one will adopt a language that has no status, in another word no benefit from speaking that language: $P_{YX}(x,0)=P_{XY}(y,0)=0$;
3. By symmetry, the transition probability should be equal when swapping the fraction of speakers and relative status: $P_{YX}(x,s_X)=P_{XY}(y,s_Y)$;

4. The rate constant c is reflect of several sociolinguistic factors, including rate of contact between pairs of individuals, the propensity for individuals to learn a new language, or political bias applied to individuals to learn a second language.
5. The relative status parameter s is the most relevant linguistically. It could serve as a useful measure of the threat to a given language. The interpretation of s will be discussed in details later in the essay together with historical data.

Plugging the expression for P_{YX} and P_{XY} from equation (2) and (3) into equation (1), one has:

$$\frac{dx}{dt} = c(1-x)s_X x^a - cx(1-s_X)(1-x)^a \qquad (4)$$

which governs the dynamics of a given language X, together with appropriate initial conditions. One may be interested in the equilibrium state of language dynamics where the number of people (x) speaking language X among the population doesn't change any more. This property can be investigated by setting dx/dt=0. Three possible solutions exist: (1) x=0, which means no one speak language X any more. (2) x=1 which corresponding to extinction of language Y. Therefore, the AS model predicts the dominance of one of the two languages and the consequent extinction of the other. This observation is independent of initial population and status. The initial condition will determine which one of the language will eventually win. The model is able to explain the decrease in the number of people speaking the endangered languages and therefore their rapid extinction, which have been shown in empirical data over last few decades. (3) Fortunately, there is a third kind of solution. For each given x, we can always find a s_X that satisfy dx/dt=0. It may not be a stable state. However, it confirmed that it is possible to have multiple competing languages coexist by controlling their status based on the fraction of population speaking each language. The tuning of the parameters of the equation can be achieved in reality by policy-making, education, and advertising to ensure the survival of the endangered language/culture.

Another important feature of AS model is to act as a alert system to tell the society when a language become endangered at an early stage, so that appropriate action can be taken to prevent the language extinction. In this aspect, one has to solve the equation (4) analytically or numerically, and then fit the solution with empirical data. Abrams and Strogatz collected data on the number of speakers of endangered language in 42 regions of Peru, Scotland, Wales, Bolivia, Ireland and Alsace-Lorraine. Four examples are shown in Fig. 1. After fitting the model solutions to the data, they obtained each parameter in equation (4) for different languages. The exponent a was found to be roughly constant across cultures, with $a=1.31\pm 0.25$. The status parameter s is particularly interesting, and could serve as a useful measure of the threat to a given language. Quechua (Fig. 1b), for example, still has many speakers in Huanuco, Peru, but its low status is driving a rapid shift to Spanish, which leads to an unfortunate situation in which a child cannot communicate with his or her grandparents.

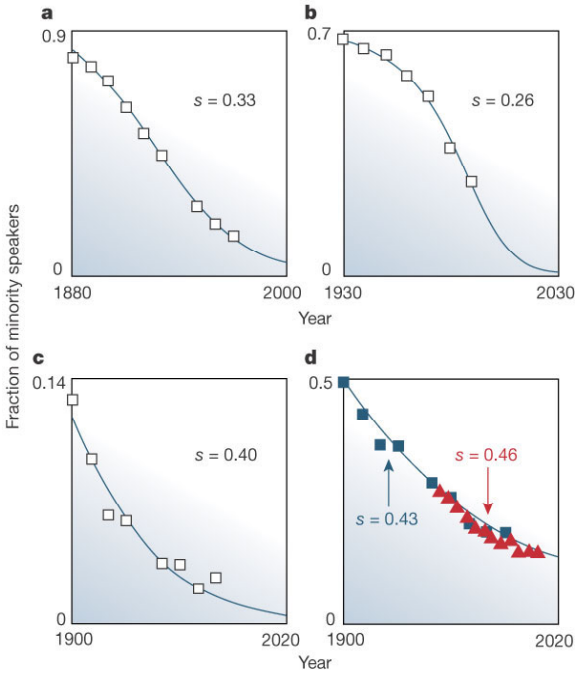

Figure 1. The dynamics of language death. Symbols show the proportions of speakers over time of: **a,** Scottish Gaelic in Sutherland, Scotland; **b,** Quechua in Huanuco, Peru; **c,** Welsh in Monmouthshire, Wales; **d,** Welsh in all of Wales, from historical data(blue) and a single modern census(red). Fitted curves show solutions of the model in equation (4), with parameters c, s, a and $x(0)$ estimated by least absolute-values regression. Where possible, data were obtained from several population censuses collected over a long timespan; otherwise, a single recent census with age-structured data was used (although errors are introduced, the size of which are reflected in the differing fits in **d**). Using the fraction of Catholic masses offered in Quechua in Peru as an indicator, we reconstructed an approximate history of the language's decline. (Figure from ref 7)

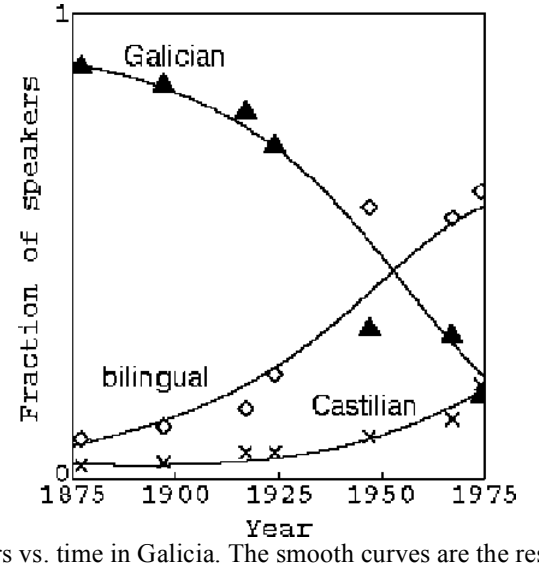

Figure 2. Fraction of speakers vs. time in Galicia. The smooth curves are the result of fitting MP model; the parameters of the fitted model are a=1.50, $s_{Galician}$=0.26, c=0.1 and k=0.80. (Figure from ref. 8)

III. MP model: modification of AS model incorporating bilingualism.

Although the AS model fit the empirical data very well, the group of people who are bilingual are ignored in the model, which is important is some cases. In addition, the prediction that bilingualism is not stable contradicts with some real world cases. Mira and Paredes modified the AS model by introducing an additional parameter to describe the similarity of two competing language [8]. This so-called MP model incorporating bilingualism push the AS model one step further and explain more data where bilingualism exists.

They generalized the AS model by introducing a bilingual group B, with fraction of population b, such that x+y+b=1. The rate equation (1) become

$$\frac{dx}{dt} = yP_{YX}(x,s_X) + bP_{BX}(x,s_X) - x[P_{XY}(y,s_Y) + P_{XB}(b,s_B)] \qquad (5)$$

where P still have the same functional form as in AS model. For example, the probability of individual transfer out from ensemble X have the form $P_{X?}=cs_?(1-x)^a$, where in AS model, the ? can be replace by Y because P_{XY} is the only possible switch. Now there are two possibilities: P_{XY} which is the possibility to become monolingual in Y and P_{XB} which is the possibility to become bilingual. Mathematically, they are

$$\begin{aligned} P_{XB}(b,s_B) &= cs_B b^a = cks_Y(1-x)^a \\ P_{XY}(y,s_Y) &= cs_Y y^a = c(1-k)s_Y(1-x)^a \end{aligned} \qquad (6)$$

where the parameter k ($0 \le k \le 1$) reflect the ease of bilingualism, in anther word the similarity of two languages. k=0 would represent no similarity, it is impossible to have conversation between X and Y, therefore $P_{XB}=0$. On the other limit, k=1 implies X=Y, where $P_{XY}=0$ and P_{XB} become as same as P_{XY} in the AS model. Similarly,

$$\begin{aligned} P_{YB}(b,s_B) &= cs_B b^a = cks_Y(1-y)^a \\ P_{YX}(x,s_X) &= cs_X x^a = c(1-k)s_X(1-y)^a \end{aligned} \qquad (7)$$

For transfer from B to X, one can assume $P_{BX}=P_{YX}$ considering that both B-to-X and Y-to-X transfers involve loss of language Y, which mainly happens due to the death of the speaker. Similarly we have $P_{BY}=P_{XY}$. Finally, we have a pair of coupled differential equation for x and y that can be solved and fit to the empirical data.

$$\begin{aligned} \frac{dx}{dt} &= c[(1-x)(1-k)s_X(1-y)^a - x(1-s_X)(1-x)^a] \\ \frac{dy}{dt} &= c[(1-y)(1-k)(1-s_X)(1-x)^a - ys_X(1-y)^a] \end{aligned} \qquad (8)$$

Again, b can be obtained by using relation x+y+b=1. Mira and Paredes fit the model to the empirical data from Galicia (northwest Spain), where Galician and Castilian coexist for more than a century. The data is shown in Fig. 2. The model fits successfully the data

and yields a high similarity between both languages (k=0.8). Further numerical calculation shows that for every s_X, there is a threshold value $k_{min}(s_X, a)$, such that when $k<k_{min}$, the language with less status will extinct over time. However for $k>k_{min}$, both group B and X survives. The conclusion of the model is that bilingualism is possible. The similarity of the competing languages is a key factor to ensure the stability of bilingualism. The MP model also allows estimating a coefficient of similarity between two languages based on historical data. It is a quantitative way to define the language boundary, which is extremely important when one try to distinguish between language and dialect. The further application in this direction will be discuss in section V.

IV. Microscopic Model

Microscopic model can take into account the internal structure of language and the interactions between individual speakers. Most of microscopic model rely on computer simulation employing Monte Carlo algorism. The most influent microscopic model is the one introduced by Schulze and Stauffer (SS model) in 2005. In this model, each language is characterized by F independent features. Each feature can take one of Q different values. The language then evolves according to three mechanisms:

1. Changes of language features. For each time step, each feature is changed with probability p. This change is random or not, depending on process 2.
2. Transfer of words from one language to another. With probability q, the change in process 1 is not random but instead transfers the value of this feature from another person in the population. With probability 1-q, the change is random.
3. The learning of a new language. With probability $(1-x)^2r$, an individual learn new language from another person in the population, where x is the fraction of people speaking the old language.

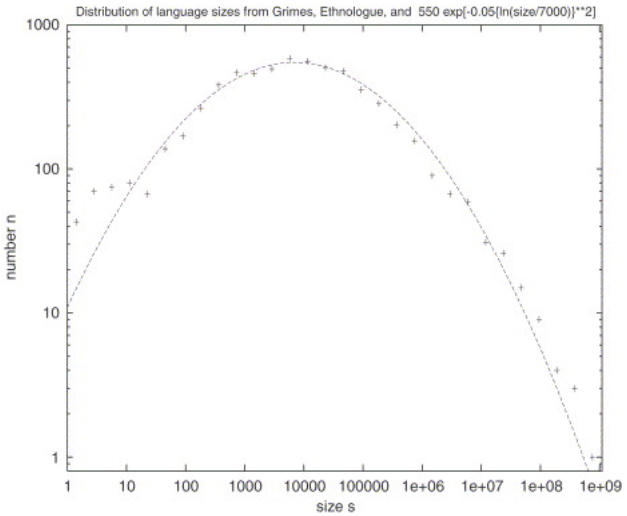

Figure 3. Distribution of language sizes (2000). The x axis represents the number of individuals speaking a language. The y axis represents the number of language spoken by x individuals. The empirical data come from Ethnologue (14^{th} editon). The dot line represents a log-normal fitting to the data. The simulation results are not included here. (Figure from ref .12)

The interaction of individuals can be models in a lattice, which each individual occupied a lattice node and only can be influenced by neighbors. More complicated network can be introduced, such as migration on the lattice. The main results from such simulations are that there is a sharp transition between one language dominates and no language dominate depending on the choice of p. The simulation also yields a log-normal distribution of language size, in agreement of empirical distribution as shown in Fig. 3.

V. Discussion

As stated by linguist Yang [14]: it is time for the ancient field of linguistics to join the quantitative world of modern science. The models I introduced in this essay are the first around attempts along the line. Although the results and conclusions from the model well explain the empirical data and look promising, many simplifying assumptions that are reasonable to physicists are not accepted very well by linguists (The same thing happened when physicist first invaded the area of biology). The key to solve this problem lies on the communication between scientists from different disciplines. The next step is to apply the models to more cases. The more successful application of the models, the more confident we feel about the models. If the models can fit the data in more cases, we can learn the threat of extinction for a given language by monitoring status parameter s in the AS model. We can also learn the similarity of two languages by computing the similarity parameter k in the MP model. If the models fail to describe the empirical data in more cases, it is an indication of huge modifications of the models are needed. In some case, the old model have to be abandoned, and new models have to be build, which is harder but more exciting. Here I propose three short-term projects that can further advance the field.

1. Case study of language extinction in China using AS model

Language extinction is a serious problem in China. The Mandarin Chinese is the official language in China with majority of speakers (845 million, which is 65% of population in China). However, there are 295 different languages currently speaking among many different ethnic groups in China. Many of them are facing extinction, as same as what is happening globally. For example, a subgroup people living in remote area of Yunnan province belong to the official minority ethnic group “Yi” speak their own language call “Ayizi”, recently identified by SIL (Summer Institute of Linguistics) international. Unfortunately, at the same time “Ayizi” being discovered, it is being listed as nearly extinct because less than 50 people are currently speaking this language. There is increasing interest in language diversity and language maintenance in China. However, the first thing need to be done is to identify the threat of extinction for a given language, which is not only depend on the population speaking the language (x in the AS model), but also the status (s in the AS model). The hard part will be collecting the empirical data over the past century for each endangered language, which may be impossible in some case due to the lack of record. Once we have the data, the AS model can help to extract information about the status, which can serve as an urgent index for language extinction in China.

2. Identification of language similarity from language competition model.

The main idea is to use MP model of language completion to quantitatively define language boundary. Ambiguity exists in the definition of a new language, especially

between chinese and western linguists. [15]. In the case of "Ayizi" mentioned above, it is recognized as a dialect by China government before 2006. Now it has been define as a new language by SIL in 2007, and listed as an endangered language. So the quantitative definition of a new language may help make the language boundary sharp and clear, which make the linguistics' life easier. The MP model offers such an opportunity. By fitting the historical data of two language and bilingualism in between, one can obtain the exact value for k, which is a measure of language similarity. In the case of Galician and Castilian studied by MP, k is 0.8, which means they are very similar to each other. It is very possible that chinese and western linguists are using different cutoff when distinguish a new language from dialect.

We can go further to verify such a speculation/hypothesis. First, we calculate k for English and Chinese from empirical data as a control experiment. I expect that k for English and Chinese should be very small because they are so different from each other. Second, we find two (or more) languages in China that spoken in the same region ensure the competition do exist between them. Third, we collect the data for two (or more) dialects in China based on same criteria. Then another set of data can be collected for western language using the same criteria in Europe, a sample will be Galician and Castilian. The k can be obtained for each set of data. It is possible that k=0.8 is small enough to be defined as two language in Europe, while in China it is classified as dialect.

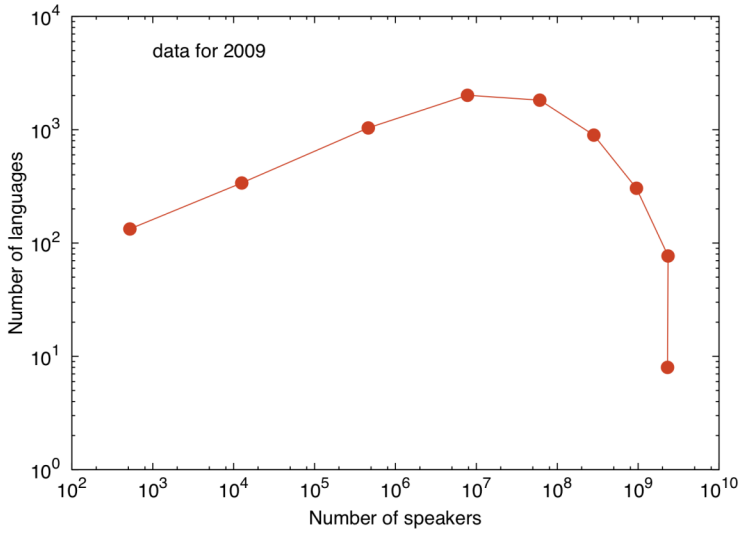

Figure 4. Distribution of language sizes (2009). The x axis represents the number of individuals speaking a language. The y axis represents the number of language spoken by x individuals which have been binned in power of ten. The empirical data come from Ethnologue (16th editon). The solid line is a simple connection of the data.

3. Time evolution of distribution of language sizes

Language evolution is dynamical. There is no reason to believe the distribution of language sizes shown in Fig. 3 is a stationary equilibrium state. I notice that the data for plotting the distribution of language sizes in the literature are based on 14th edition of "Ethnologue: Languages of the World" published in 2000. So I find the latest data

published in 2009 and plot it in Fig. 4. Surprisingly, the shape and position of the peak of the distribution both changed dramatically over the last ten years. The position of the peak language size have increased from the language with ten thousands speakers to ten million speakers. The increase is on the order of three magnitudes. It may be interesting to look at the historical data for language sizes distribution. Its evolution can be modeling as well. The shifting of the peak can be attribute to the rapid globalization process over the past 20 years. More and more people speak the same language, such that small language clustering into a common language spoken by more people. It is also an indication of language extinction we are facing, as well as the increase in the extinction rate, which are interesting for further investigations.

References:

1. Lewis, M. Paul (ed.), (2009) Ethnologue: Languages of the World, Sixteenth edition. Dallas, Tex.: SIL International. www.ethnologue.org
2. Krauss, M. (1992) Endangered languages- the world language in crisis. Language **68**, 4-10
3. Crystal, D. (2000) Language death, Cambridge University Press, Cambridge
4. Sutherland, W. J. (2003) Parallel extinction risk and global distribution of languages and species. Nature. **423**, 276-279
5. Castellano, C., Fortunato, S., and Loreto, V. (2009) Statistical physics of social dynamics. Reviews of Modern Physics. **81**, 591-646
6. Loreto, V., and Steels, L. (2007) Emergence of language. Nature physics. **3**, 758-760
7. Abrams, D. M., and Strogatz, S. H. (2003) Modeling the dynamics of language death. Nature. **424**, 900
8. Mira, J., and Paredes, A. (2005) Interlinguistic similarity and language death dynamics. Europhysics letters. **69**, 1031-1034
9. Wang, W. S-Y., and Minett, J. W. (2005) The invasion of language: emergence change and death. Trends in Ecology and Evolution. 20, 263-269
10. Minett, J. W., and Wang, W. S-Y. (2008) Modeling endangered languages: the effects of bilingualism and social structrue. Lingua, **118**, 19-45
11. Stauffer, D., Schulze, C., Lima, F. W. S., Wichmann S., and Solomon, S. (2006) Non-equilibrium and irreversible simulation of competition among languages. Physica A, **371**, 719-724
12. Schulze, C., Stauffer, D., and Wichmann S. (2008) Birth, Survival and Death of Languages by Monte Carlo Simulation. Communications in computational physics. **3**, 271-294
13. de Oliveria, V. M., Gomes, M. A. F., Tsang, I. R. (2006) Theoretical model for the evolution of the linguistic diversity. Physica A, **361**, 361-370
14. Yang, C. (2006) The infinite gift-how children learn and unlearn the language of the world, Scribner, New York.
15. Erard, M. (2009) How many languages? Linguists discover new tongues in China. Science, **324**, 332-333