Physica A 388 (2009) 2874–2879

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A simple branching model that reproduces language family and language population distributions

Veit Schwämmle [a,b,*], Paulo Murilo Castro de Oliveira [a,c,d]

[a] Laboratoire PMMH, École Supérieure de Physique et de Chimie Industrielles, 10 rue Vauquelin, F-75231 Paris, France
[b] Centro Brasileiro de Pesquisas Físicas, Rua Xavier Sigaud 150, 22290-180 Rio de Janeiro, Brazil
[c] Instituto de Física, Av. Litorânea s/n, Boa Viagem, Niterói 24210-340, RJ, Brazil
[d] National Institute of Science and Technology for Complex Systems, Brazil

ARTICLE INFO

Article history:
Received 14 November 2008
Received in revised form 16 March 2009
Available online 28 March 2009

Keywords:
Complex system
Branching process
Population distributions
Language classification

ABSTRACT

Human history leaves fingerprints in human languages. Little is known about language evolution and its study is of great importance. Here we construct a simple stochastic model and compare its results to statistical data of real languages. The model is based on the recent finding that language changes occur independently of the population size. We find agreement with the data additionally assuming that languages may be distinguished by having at least one among a finite, small number of different features. This finite set is also used in order to define the distance between two languages, similarly to linguistics tradition since Swadesh.

1. Introduction

The existence of the large number of around 6000 languages on Earth can be explained through their continuous modification. They descend through a tree-like evolution from one or a few proto-languages several thousand years ago. Language evolution is thus a result of the particular history followed by humankind and its migration pattern. On the other hand, human genetic evolution is also a result of the same history. The parallel between language and genetic evolution was explored by many researchers since Cavalli-Sforza (for a review, see Ref. [1]), in order to discover unknown details of the human past history.

In addition to the present situation of spoken languages, we are also interested in their historical course, resulting from the branching of ancestor languages. Branching models have already been used to model biological systems for a longer time (e.g., see Ref. [2]). The direct comparison of selected word sets of different languages can be used to estimate their historical distance, an idea pioneered by Morris Swadesh half a century ago [3]. For instance, the measurement of Levenshtein distances between two languages gives an idea of the time their first common ancestor language existed. The evaluation of the data from this analysis showed that historical distances accumulate at a certain age independently of the population size [4,5], suggesting that change occurs roughly with the same rate for all languages [6]. This result challenges the often proposed direct analogy between biological and language evolution and demands different approaches. In contrast, the global mutation rate of biological species depends on their population size, leading for instance to faster changes of the genetic pool in smaller populations.

The statistical analysis of highly complex systems like opinion dynamics and stock exchange among others showed that their patterns can be reproduced by simple agent-based models. The simplification of most of the complex low-level

* Corresponding author at: Laboratoire PMMH, École Supérieure de Physique et de Chimie Industrielles, 10 rue Vauquelin, F-75231 Paris, France. Tel.: +55 21 26295829; fax: +55 21 26295887.
E-mail address: veitveit@gmail.com (V. Schwämmle).

0378-4371/$ – see front matter
doi:10.1016/j.physa.2009.03.038

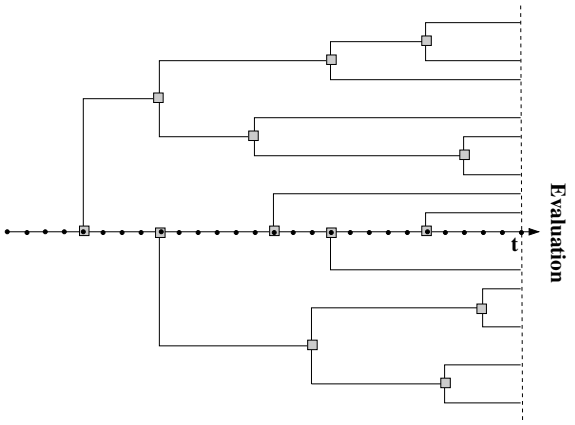

Fig. 1. Schematic description of the language model. Mutations occur according to the same rate for all languages, at the time steps visualized by small black dots along the time axis. At each time step, a branching event occurs with probability b, visualized by grey squares. After reaching a previously defined number of branches, vertical line on the right, the simulation stops and the data are evaluated.

mechanisms to random processes has been shown to be a valid approach. We follow this idea and will construct a simple model that uses the main properties found for the evolution of languages. Our stochastic model is based on the following assumptions: (i) Languages evolve in a tree-like structure. (ii) The structure is modeled by having a probabilistic change rate that does not depend on the language's number of speakers. (iii) The space of possible languages is finite, the same language can be visited from different evolutionary paths. (iv) Each population exponentially increases in time.

Assumption (iii) needs further explanation: The method to compare languages relies on the comparison of word sets. This procedure is based on the two following assumptions. First, it is sufficient to take the language features of a small number of native speakers to represent the entire language. We neglect small variations between different speakers. Second, only a small characteristic word set is chosen for comparison. Therefore, by further refining the method, for instance comparing a larger number of speakers and taking larger word sets, the total number of different languages increases until obtaining the result that everybody speaks a different one. On the other side, a coarsening would lead to the limit that only one common language rules Earth. Hence, by allowing small variations and comparing an accurately chosen set of words, we are able to distinguish a large but not too large number of languages. However, we need also to remind that the number of possible different languages has a finite upper limit with this method.

Motivated by the increasing popularity of simulating language evolution and competition, several analytical and computational models concentrated on reproducing the histogram of language sizes. For instance, its lognormal-like shape may be explained by simply assuming independent language population growth. These independent processes naturally lead to a lognormal distribution being result of the central limit theorem for multiplied random variables [7]. They can also be coupled with a fragmentation process to yield distributions from lognormal distributions to power-laws [8]. However, the distribution of real languages presents a deviation from lognormal behavior for languages spoken by very few people. Agent-based models found agreement with the real data furthermore reproducing the deviation for small population sizes. This was achieved by measuring the histogram in a simulation state before reaching the absorbing state [9], or by allowing redundancy in creating new languages, i.e. by allowing different historical paths leading to the same final language [10,11]. The latter model also considers the gradual geographical conquest of new territories, introduced in Refs. [12,13], and until now may be seen as the most appropriate one as it not only reproduces the histogram of language sizes but also the distribution of languages' family sizes recently reported [14]. Another recent work proposes that the histogram of language sizes deviates from the lognormal shape also for large populations displaying power-law decays at both extremes [15]. For a thorough review of language models see Refs. [16,17]. We try here to gather the principal findings from previous models in order to construct a yet simpler model with a reduced number of parameters that reproduces the real data sets.

This work is organized as follows. The following section introduces the model. The next section presents our results and their direct comparison to real data. Finally, we conclude and discuss our findings in the last section.

2. Branching model

Our model is agent-based, i.e. each language has its own characteristic traits. A language is characterized by its number of speakers, given by a real number, its structure condensed into a bit-string of L features, and an integer number defining the language family it makes part of. Time evolution takes place in discrete time steps having the following rules applied to each language present at a given time step. (i) The number of speakers is multiplied with the factor $1+G$ leading to its exponential increase. (ii) The current bit-string characterizing the language has one of its bits flipped at a random position, from 0 to 1 or vice-versa. (iii) The grown language divides into two with probability b. In the case of branching, three additional processes occur. First, the population is distributed by giving a randomly chosen part between 0% and 50% of the ancestor population to the new branching language and leaving the remaining population to the ancestor. Afterwards, the bit-string

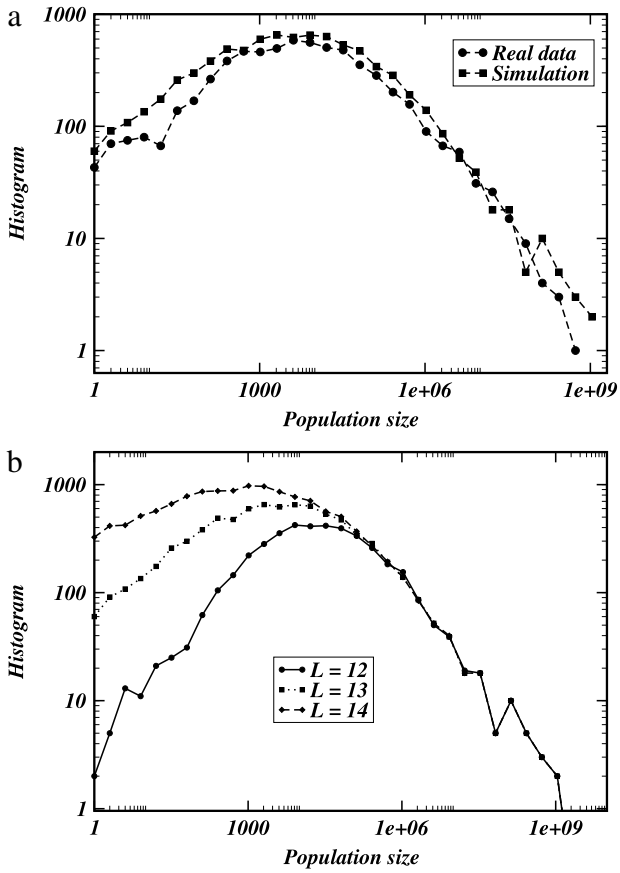

Fig. 2. (a): Comparing the histograms of language populations: real data in the world *versus* simulations. The vertical axis counts the number of languages spoken by the number of people displayed along the horizontal axis. (b): Comparing the histogram for different L.

of the new language suffers a flip of one of its bits, compared with the ancestor, causing them to differ from each other by one feature. Third, if the new language still belongs to the first family, it founds a new language family with probability $F = 0.5$. Otherwise it obtains the family label from the ancestor.

The simulation initializes with one single language having all bits set to 0. As there are no extinctions, the total number of branches increases exponentially with time until we stop the simulation after reaching the previously given threshold of $N_{\mathrm{L,max}}$ branches. We could easily consider also language extinctions with a given probability, but this would introduce a further, unnecessary parameter into the model: the absence of some branches simply corresponds to a smaller value of the branching parameter.

After the final time step we process the data, consisting of a large number of branches, each having its corresponding bit-string, its number of speakers and the family it pertains to. The data evaluation defines which branches differ, and which are bundled to represent only one language. More precisely, we compare the bit-strings of each pair. If the language structures are identical, both populations are added to form a unified language. Additionally, if their family labels, assigned during the previous dynamic evolution, differ from each other, the family of the one with the larger population is kept. Depending on the number of language features, i.e. bit-string length L, the number of languages after evaluation will be less or equal to 2^L. Fig. 1 shows a scheme of our model.

3. Comparing the results to real data

As a classification of languages, we take the database of Ethnologue which is accessible through the internet [18]. This also provides the number of speakers of a language, defined by the number of people speaking a language as their native language. These data will be compared to the simulational results.

Both real and simulational data present ranges of population sizes over about nine decades. Therefore it is convenient to plot the histogram in the following way: we count the number of languages in a bin with population sizes between 2^n and 2^{n+1}. As a consequence, the size of the bins increases exponentially. We plot these numbers without dividing them by the length of the bin. Therefore, they do *not* correspond to the frequency of languages with population sizes within the given ranges.

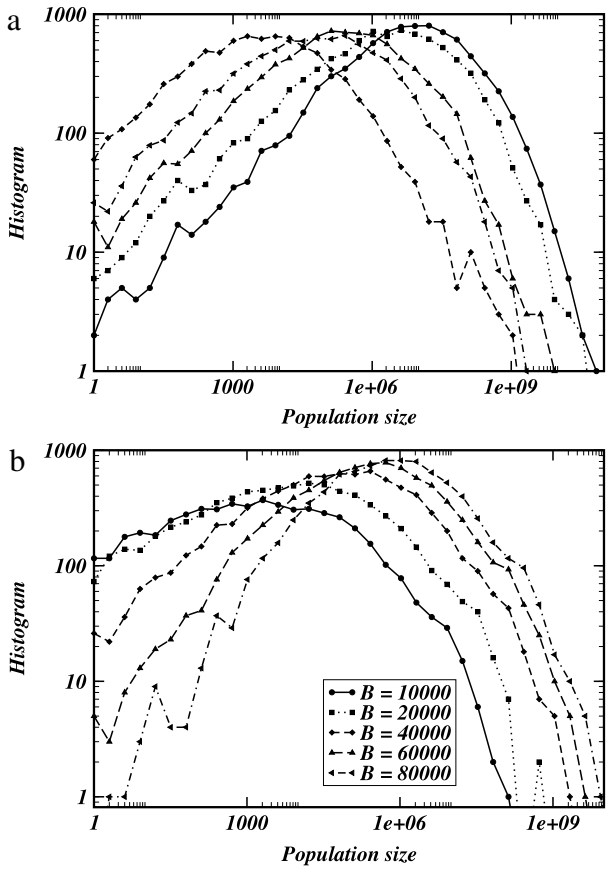

Fig. 3. (a): The histogram of language sizes for different random seeds leads to quantitatively different results. (b): The distribution shifts to larger populations and larger language numbers when increasing final number of branches, B, equivalent to an increase of the simulation time. Simulations in both figures have been carried out with the same parameter set as in Fig. 2.

By making a histogram of language population sizes (Fig. 2a), we found good agreement between simulational results and real data for a bit-string length of $L = 13$, a branching probability $b = 0.01$, and a growth rate $G = 0.023$. The simulation stopped as soon as the tree had more than 40,000 different branches, corresponding to 1024 time steps and a total population of 1.3×10^{10} speakers. As we defined the languages to be different only if they exhibit distinct features, the final number of languages was 8125, near the maximum value allowed with $L = 13$, indicating the possibility of re-visiting the same language by different historical paths. On the other hand, within similar numbers of languages but with larger values of L, the possibility of re-visiting the same language more than once becomes negligible. In this case, our simulations also result in lognormal-like distributions as in Fig. 2(a), but with a stronger deviation observed at the very left part, for languages spoken by very few people (Fig. 2(b)). Based on these observations, we can tentatively interpret these deviations: languages spoken by very few people, separated in small groups which nevertheless exchange some experiences with each other, may be the results of different historical paths leading to the same final language, i.e. historical redundancy. This multiple-historical path for the same language indeed also occurs for large languages (spoken by millions of people). For instance, different regions of the same large country have small linguistic differences, which were not captured by the linguistics finite set of words adopted in order to distinguish one language from another. This leads to the decrease of counted languages with small populations. According to this interpretation, the quoted deviations would become more accentuated if one is able to enlarge the set of words adopted nowadays in order to distinguish languages, i.e. by enhancing the "resolution". Yet according to this interpretation, the resolution currently adopted by professional linguists is just enough to distinguish the main real human languages from each other. The measuring instrument is designed with the resolution degree one needs for the object at hands, no more.

The simulational results strongly depend on the random seed set at the beginning of the simulation in order to generate the sequence of pseudo-random numbers. The randomly chosen time steps at which the first branching events occur are crucial, and can lead to different simulational results (Fig. 3(a)). The same behavior is also expected in reality, the real historical path depends on contingencies which have occurred during the evolution, and this dependence is supposed to be stronger for contingencies which have occurred at the very beginning. Unfortunately, one is not able to re-run the real evolution again, only computer simulations like the current one allow this repetition, by taking different random seeds. By

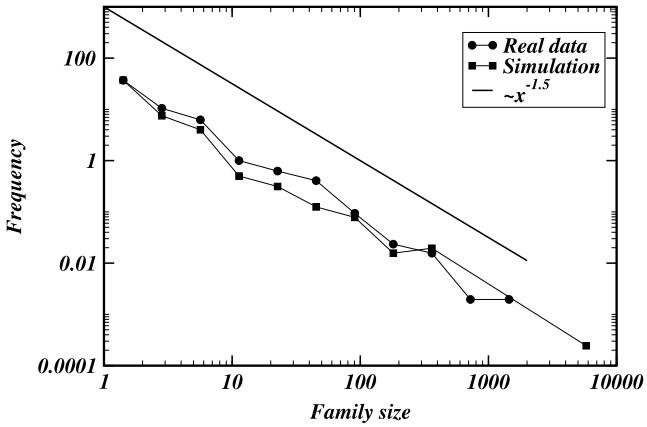

Fig. 4. Frequency of language families comparing simulational and real data. The same power-law is found in both systems.

doing so, the curve of the histogram of population sizes moves towards larger or smaller populations. Thus different random seeds result in a horizontal shift on the histogram.

By changing the total number of time steps we are additionally able to perform vertical shifts provided that the number of languages is still sufficiently below its maximum value. However, in any case the lognormal shape of the distribution, including the deviation for languages spoken by few people remains the same. Histograms with smaller widths and smaller heights could be interpreted as describing early stages of human language evolution.

The two parameters, branching probability b and growth rate G, have a similar effect. By increasing the branching probability, the number of time steps needed to reach the previously fixed maximum number of branches is smaller, and therefore the final total population becomes smaller as it grows exponentially with time. The shape of the histogram remains unaltered. Nevertheless, the strong dependence on the random seed makes a scaling of this parameter difficult. On the other side, the growth rate has the opposite effect on the simulational results. By linearly increasing it, the total population with fixed number of branches at the end of the simulations increases exponentially.

Finally, for language numbers near the maximum value, the distribution shifts to larger population sizes for longer simulation times as could been expected. Additionally, it becomes more narrow as a maximum number of languages is predetermined through the bit-string length 13 (Fig. 3(b)).

The simulations presented here are also valid for language subgroups, i.e. by taking an arbitrary branch of the tree as the origin of smaller tree we obtain the same results.

Now we look to see if our model also reproduces the other simple macroscopic law that has recently been reported for languages [14]. The size distribution of language families displays a power-law decay. The size of a family is obtained by counting its languages. In Fig. 4, we show the frequency of the families, again for a binning over powers of 2. We count the number of families with sizes between 2^n and 2^{n+1} and divide this number by the length of the binning, 2^n. The simulational data and the real data agree nicely with each other, exhibiting a power-law with an exponent of about -1.5. The parameter values are the same as the ones for the simulation shown in Fig. 2. Note, that as soon as a branch belongs to a language family, it is conserved during the simulation. Therefore, in this case language subgroups do not yield the same results.

4. Conclusions

Based on a few basic assumptions, we constructed a simple branching model and compared its statistical results with real data of human languages and families of languages. The good agreement shows that our model may be considered a minimal model for the evolution of languages. In particular, we need not consider any geographical issue.

The only important assumptions we need are: (i) All languages descend from a single, ancient mother-tongue; (ii) All languages change according to the same and constant rate, independent of the number of speakers; (iii) Bifurcation events (a single language leading to two different ones) occur according to a smaller rate, also constant and the same for all languages; (iv) The different features allowing us to distinguish two languages from each other can be condensed into a finite, small set; (v) The founder of a new language family always belongs to the original family founded by mother-tongue. The last point does not mean that one cannot go further into the taxonomic classification, by taking into account language genera, etc. We simply did not treat this issue within this work.

Finally, we predicted how the histogram of language sizes changes when refining or coarsening the classification method.

References

[1] L.L. Cavalli-Sforza, Genes, Peuples et Langues, Odile Jacob, Paris, 1996.
[2] J. Chu, C. Adami, Proc. Natl. Acad. Sci. 96 (1999) 15017.

[3] M. Swadesh, Int. J. Am. Linguistics 21 (1955) 121.
[4] S. Wichmann, D. Stauffer, E.W. Holman, Adv. Complex Syst. 11 (2008) 357.
[5] F. Petroni, M. Serva, J. Stat. Mech. (2008) P08012.
[6] S. Wichmann, P.M.C. de Oliveira, V. Velupillai, A. Müller, D. Bakker, A. Grant, A universal law of language taxonomics, 2008, preprint.
[7] D.H. Zanette, Int. J. Mod. Phys. C 19 (2008) 237.
[8] Ç. Tuncay, Int. J. Mod. Phys. C 19 (2008) 471.
[9] D. Stauffer, C. Schulze, F.W.S. Lima, S. Wichmann, S. Solomon, Physica A 371 (2006) 719.
[10] P.M.C. de Oliveira, D. Stauffer, F.W.S. Lima, A.O. Sousa, C. Schulze, S. Moss de Oliveira, Physica A 376 (2007) 609.
[11] P.M.C. de Oliveira, D. Stauffer, S. Wichmann, S. Moss de Oliveira, J. Linguistics 44 (2008) 659.
[12] V.M. de Oliveira, M.A.F. Gomes, I.R. Tsang, Physica A 361 (2006) 361.
[13] V.M. de Oliveira, P.R.A. Campos, M.A.F. Gomes, I.R. Tsang, Physica A 368 (2006) 257.
[14] S. Wichmann, J. Linguistics 41 (2005) 117.
[15] V. Schwämmle, S.M.D. Queirós, E. Brigatti, T. Tchumachenko, Competition and fragmentation: A simple model generating lognormal-like distributions, 2008, arXiv.org:0810.2403.
[16] C. Schulze, D. Stauffer, S. Wichmann, Comm. Comp. Phys. 3 (2008) 371–394. arxiv:0704.0691.
[17] S. Wichmann, D. Stauffer, F.W.S. Lima, C. Schulze, Trans. Phil. Soc. 105 (2007) 126.
[18] B.F. Grimes, Ethnologue: Languages of the World, 14 edition, Summer Institute of Linguistics, Dallas, TX, 2000, www.sil.org.