Bounded fitness landscapes and the evolution of the linguistic diversity

Viviane M. de Oliveira[a]*, Paulo R. A. Campos[b], M. A. F. Gomes[a], I. R. Tsang[c]

June 3, 2006

[a]Departamento de Física, Universidade Federal de Pernambuco, 50670-901, Recife, PE, Brazil
[b]Departamento de Física e Matemática, Universidade Federal Rural de Pernambuco 52171-900, Dois Irmãos, Recife-PE, Brazil
[c]Centro de Informática, Universidade Federal de Pernambuco, 50670-901, Recife, PE, Brazil

Abstract

A simple spatial computer simulation model was recently introduced to study the evolution of the linguistic diversity [1]. The model considers processes of selective geographic colonization, linguistic anomalous diffusion and mutation. In the approach, we ascribe to each language a fitness function which depends on the number of people that speak that language. Here we extend the aforementioned model to examine the role of saturation of the fitness on the language dynamics. We found that the dependence of the linguistic diversity on the area after colonization displays a power law regime with a non-trivial exponent in very good agreement with the measured exponent associated with the actual distribution of languages on the Earth.

*viviane@df.ufpe.br

1 Introduction

The research in language dynamics has arose an increasing interest of the complex systems community in the last years. Most of the researchers focus their investigations on issues like rise, competition, extinction risk and death of languages [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Furthermore, recent advances in archeology, genetics and linguistics have provided relevant contributions to a better comprehension of the linguistic diversification [12, 13]. Some investigations have demonstrated that distinct causes have greatly affected the evolution of the linguistic diversity. Among the main elements are geographic factors, economic features, complexity of the language, to cite just a few. For instance, Sutherland [2] has shown that beside country area, forest area and maximum altitude contribute to increase diversity, whereas the diversity decreases for a larger latitude. According to Bellwood [14, 15] and Renfrew [16, 17] the occurrence of agricultural expansion was the responsible for the massive population replacements initiated about 10,000 years ago and caused the disappearance of many of the Old World languages.

In a recent work, we investigated the evolution of the linguistic diversity by introducing a spatial computer simulation model that considers a diffusive process which is able to generate and sustain the diversity [1]. The model describes the occupation of a given area by populations speaking several languages. To each language was assigned a fitness value f which is proportional to the number of sites colonized by populations that speak that language. In the process of colonization, language mutation or differentiation and language substitution can take place, which affords the linguistic diversity. This simple model gives rise to scaling laws in close resemblance with those reported in [18].

In the current contribution, we study the dynamics of the linguistic diversity but now we assume that the fitness of each language is bounded by a given maximum (saturation) value which is randomly chosen from an uniform distribution. The saturation hypothesis mimics factors like the difficulty/ease of learning the languages and economy that permit some languages to propagate more easily than others.

The paper is organized as follows. In Section 2 we introduce the model. In Section 3 we discuss the results. And finally, in Section 4 we present the conclusions.

2 Model

Our model is defined on a two-dimensional lattice of linear size L, and composed of $A = L \times L$ sites with periodic boundary conditions. Each lattice site s_i represents a given region, which can be occupied by a single population speaking just one language. We ascribe to each site a given capability C_i, whose value we estimate from a uniform distribution, defined in the interval 0-1. The capability means the amount of resources available to the population which will colonize that place. It is implicit that the population size in each cell s_i is proportional to its capability C_i.

In the first step of the dynamics, we randomly choose one site of the lattice to be colonized by a single population that speaks the ancestor language. Each language is labeled by an integer number. As soon as a new language arises, it is labeled by the next upper integer. To each language, we assign a fitness value f, which is calculated as the sum of the capabilities of the sites which speak that specific language. But now differently from reference [1], the fitness can not exceed an integer value γ_k which we have chosen to be in the range 1-2000. This saturation term γ_k is randomly chosen when the language k appears. Thus, the initial fitness of the ancestor language is the capability of the initial site.

In the second step, one of the four nearest neighbors of the site containing the ancestor language will be chosen to be colonized with probability proportional to its capability. We assume that regions containing larger amount of resources are most likely to be colonized faster than poor regions. The referred site is then occupied by a population speaking the ancestor language or a mutant version of it. Mutations are the mechanisms responsible for generating diversity, and together with the natural selection mantains the standing level of diversity on the system. The probability of occurrence of a mutation in the process of propagation is $p = \frac{\alpha}{f}$, where α is a constant, and so the mutation probability is inversely proportional to the fitness of the language. The form of the mutation probability p is inspired by population genetics, where the most adapted organisms are less likely to mutate than poorly adapted organisms [19]. The probability of producing reverse mutations is zero, that is, the language generated by a mutation is always different of the previous ones.

In the subsequent steps, we check the empty sites which are located on the boundary of the colonized cluster, and we then choose one of those empty sites according to their capabilities. Again, those sites with higher capabili-

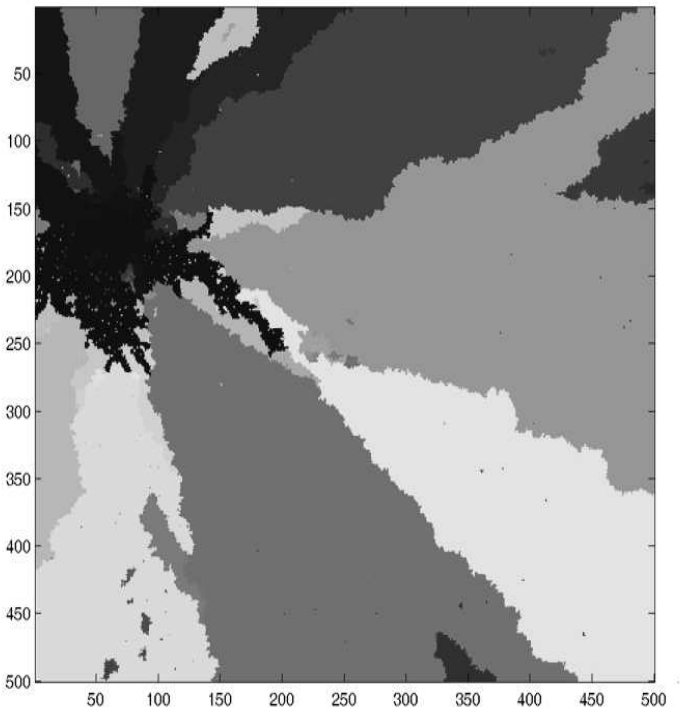

Figure 1: Snapshot of a typical realization of the dynamics at the first moment of colonization of all sites. The saturation quantities γ_k are randomly chosen in the interval 1-2000. The lattice size is $L = 500$ and $\alpha = 0.3$. See text for detail.

ties enjoy of a greater likelihood to be occupied. After that, we choose the language to be incorporated in the chosen cell among those languages occupying the neighboring sites. Languages with higher fitness have higher chance to expand. The process continues while there are empty sites in the network. After completion, we count the total number of languages D. In order to give to the reader some insight about our model, in Figure 1 we present the snapshot for a typical realization of the dynamics at the first moment of colonization of all sites (in this figure the gray scale represents different languages). The striated linguistic domains presenting very small territories occupied by different languages shown in Figure 1 remind us the actual distribution of languages observed in the Caucasus region between Black and Caspian Seas, a relatively small area of 300,000 km^2 where languages of the Caucasic, Indo-European and Altaic families coexist distributed within a large variety of peoples [13].

3 Results and Discussion

In Figure 2, we show the diversity D as a function of the area A (total number of sites in the lattice) for mutation parameter $\alpha = 0.3$ and saturation values defined in the interval 1-2000. The points are averages over 100 independent simulations when $L < 400$ and over 20 simulations when $L = 500$. We observe that the curve presents just one scaling region which extends over five decades. The exponent $z = 0.39 \pm 0.01$ is in quite satisfactory agreement with the exponent observed for the actual distribution of languages on Earth. For sake of completeness, we also exhibit in Figure 2 the observed values ($*$) of diversity versus area obtained in reference [18] for all languages spoken on Earth (the ten data points are associated with the interval from $A = 50$ km^2 to $A = 10^7$ km^2 of the actual distribution). We notice in passing that although there is not a perfect scaling relationship between diversity and area along five decades in area, both the simulation and the actual data of $D(A)$ curiously seem to be modulated by a similar tendency to oscillate in respect to the main scaling behavior (the deviations from perfect scaling in the actual data have no connection with the choice of the bins). We have also investigated the situation at which the saturation value is the same for all languages. We have noticed a linear growth of the diversity with area when the maximum γ is very small. For large values of γ we notice the existence of two scaling regions. For very large values of γ we recover the result obtained

for the case where the fitness are not limited [1].

Figure 3 displays the number of languages with population size greater than N, $n(> N)$, as a function of N. In order to obtain the curves, we have assumed that the population in a given site is proportional to the capability in the site. We have considered that the population in a given site is its capability multiplied by a factor 100. In the plot, the values of the parameters are $L = 500$ and $\alpha = 0.3$. In close analogy with the distribution of languages on Earth [18], we find two distinct scaling regimes $n(> N) \sim N^{-\tau}$: $\tau = 0.35 \pm 0.01$ for $200 < N < 2,000,000$, and $\tau = 1.14 \pm 0.01$ for $2,000,000 < N < 10,000,000$. The inset exhibits the differential distribution of languages spoken by a population of size N, $n(N)$. This distribution also agrees with the one observed for languages on Earth [18, 2]; in particular it is well described by the lognormal function $n(N) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{N}\exp\left[-\frac{1}{2\sigma^2}(\log N - \mu)^2\right]$, with $\sigma = 0.41$ and $\mu = 0.42$ (continuous curve in the inset).

4 Conclusions

We have introduced a model for evolution of linguistic diversity that considers a bounded fitness value for languages. We have considered a random chosen value of saturation of the fitness for each language in order to mimic the fact that different languages have different conditions to propagate. We have noticed a considerable improvement of the results when compared to the earlier approach [1]. Now, the relationship between diversity and area presents just one scaling regime. For $\alpha = 0.3$ we obtain $z = 0.39 \pm 0.01$, which is in very good agreement with the exponent observed for the languages on the Earth [18], along five decades of variability in area. We have also observed that the exponents τ for the two power law regimes in $n(> N)$ as a function of N are closer to those obtained by empirical observations [18].

In order to compare other kinds of saturation conditions, we have also studied the case where the saturation values are the same for all the languages. With this condition, we could not reproduce the basic relationship between diversity and area observed for the actual distribution of languages, although for the very particular and unrealistic case where $\alpha = 0.01$ and $\gamma = 1$, we can perfectly reproduce the differential distribution of languages spoken by a population of size N, $n(N)$, as well as the number of languages with population size greater than N, $n(> N)$, as a function of N. Our results seem to demonstrate that different assumptions on the behavior of the

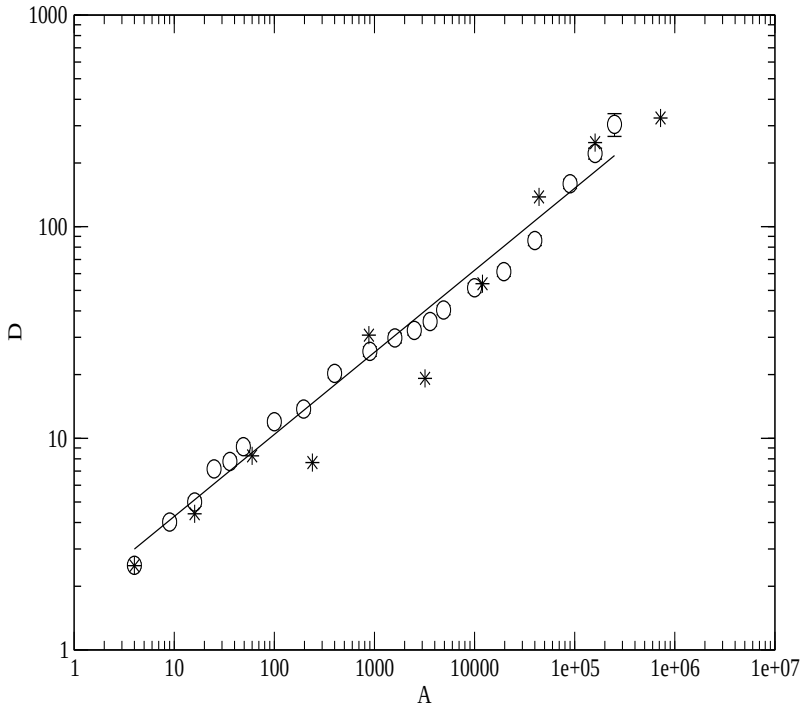

Figure 2: Number of languages D as a function of the area A for $\alpha = 0.3$. The exponent is $z = 0.39 \pm 0.01$. The asterisks represent data from the actual distribution of languages on Earth. See text and Figure 1 of reference [18] for detail.

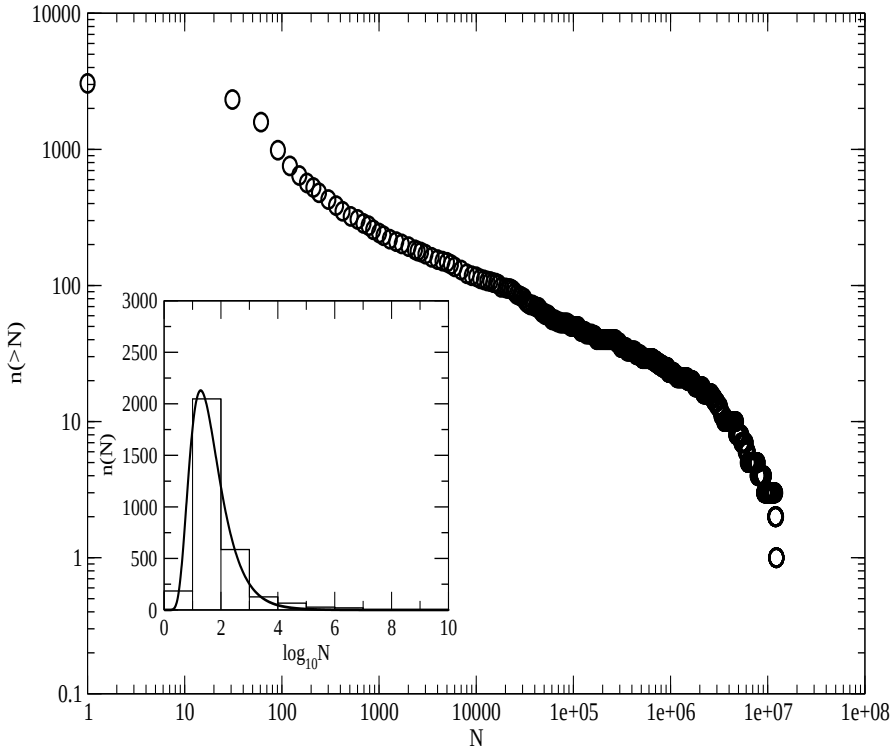

Figure 3: Main plot - number of languages with population greater than N, $n(> N)$, as a function of N. $n(> N) \sim N^{-\tau}$ with $\tau = 0.35 \pm 0.01$ for $200 < N < 2,000,000$ and $\tau = 1.14 \pm 0.01$ for $2,000,000 < N < 10,000,000$. Inset - corresponding differential distribution $n(N)$ with lognormal best fit (continuous line). See text for detail.

fitness function have very important consequences on the characteristics of the language spreading.

V. M. de Oliveira and M. A. F. Gomes are supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico and Programa de Núcleos de Excelência (Brazilian Agencies). P. R. A. Campos is supported by CNPq.

References

[1] V. M. de Oliveira, M. A. F. Gomes, I. R. Tsang, Physica A, in press.

[2] W. J. Sutherland, Nature **423** (2003) 276.

[3] D. M. Abrams, S. H. Strogatz, Nature **424** (2003) 900.

[4] M. Patriarca, T. Leppänen, Physica A **338** (2004) 296.

[5] C. Schulze and D. Stauffer, Int. J. Mod. Phys. C **16**(5) (2005) 781.

[6] J. Mira, A. Paredes, Europhys. Lett. **69** (2005) 1031.

[7] V. Schwämmle, Int. J. Mod. Phys. C **16**(10) (2005)

[8] K. Kosmidis, J. M. Halley, P. Argyrakis, Physica A **353** (2005) 595.

[9] D. Stauffer, C. Schulze, Phys. of Life Rev. **2** (2005) 89.

[10] T. Tesileanu, H. Meyer-Ortmanns, arXiv:physics/0508229

[11] J. M. Diamond, Nature **389** (1997) 544.

[12] C. Renfrew, Man **27**, No. 3, (1992) 445.

[13] L. L. Cavalli-Sforza, Genes, Peoples and Languages, Penguin, London, 2001.

[14] P. Bellwood, in The Origins and Spread of Agriculture and Pastoralism in Eurasia (ed. Harris, D. R) UCL Press, London, 1996, pp. 465-498.

[15] P. Bellwood, Sci. Am. **265**(1) (1991) 88.

[16] C. Renfrew, in The Origins and Spread of Agriculture and Pastoralism in Eurasia (ed. Harris, D. R) ,UCL Press, London, 1996, pp. 70-92.

[17] C. Renfrew, Archaeology and Language: The Puzzle of Indo-European Origins, Cape, London, 1987.

[18] M. A. F. Gomes, G. L. Vasconcelos, I. J. Tsang, I. R. Tsang, Physica A **271** (1999) 489.

[19] N. H. Barton, in: D. Otte, J. A. Endler (Eds.), Speciation and Its Consequences, Sinauer Associates, Sunderland, MA, 1989, pp. 229-256.