Scaling Relations for Diversity of Languages

M.A.F.Gomes[1], G.L.Vasconcelos[1], I.J.Tsang[2], and I.R.Tsang[2]

[1] *Departamento de Física, Universidade Federal de Pernambuco, 50670-901 Recife, PE, Brazil*

[2] *Department of Physics, University of Antwerp, Antwerp B-2020 Belgium*

(January 8, 2006)

Abstract

The distribution of living languages is investigated and scaling relations are found for the diversity of languages as a function of country area and population. These results are compared with data from Ecology and from computer simulations of fragmentation dynamics where similar scalings appear. The language size distribution is also studied and shown to display two scaling regions: (i) one for the largest (in population) languages and (ii) another one for intermediate-size languages. It is then argued that these two classes of languages may have distinct growth dynamics, being distributed on sets of different fractal dimensions.

PACS numbers: 05.40.+j, 64.90.+b, 89.60.+x, 89.90.+n

Keywords: Diversity, Languages, Fragmentation, Fractals

I. INTRODUCTION

A great deal of effort has been made to know the Earth's biodiversity [1]. In spite of this, only about 1.7 million of an estimated 13.6 million species have been identified to date. The diversity of languages, on the other hand, is much better known: there are 228 countries in the world with a total of approximately 5000 ethnic groups speaking about 6500 different languages [2,3]. In this Paper we report a quantitative analysis of how language diversity increases with country area and population. A study of the language size distribution is also presented.

The concept of diversity plays an important rôle in an increasing number of contexts in the scientific literature in connection with biological problems [4], cellular automata [5], diffusion processes [6], ecological [7] and evolutionary [8] problems, fractals [9], and fragmentation phenomena in general [10], including nuclear fragmentation [11]. Scaling relationships between diversity and the system size has been reported in a number of studies on fragmentation [6,10,12,13] and Ecology [7,14–16]. For example, it has now been firmly established, from both ecological field data [14] and computer models [7,15], that the number of species or diversity D in a ecosystem of area A increases with A as a power law: $D \sim A^z$, where the exponent z varies from 0.1 to 0.45 [14,15].

Here we report a scaling relationship between language diversity and area that is akin to the relation above observed in Ecology. A scaling relation has also been found between language diversity and population. The diversity distribution, meaning the number of countries with a given diversity, displays composite power-laws, and an argument is presented to account for the existence of these two distinct scaling regimes. We also study the language size distribution and show that it has two scaling regions corresponding, respectively, to (i) the largest (in population) languages and (ii) intermediate-sized languages. It is then argued that the existence of distinct scaling behaviors for these two classes of languages imply that they have distinct growth dynamics, leading to different patterns of space occupation, such as different fractal dimensions.

II. RESULTS AND DISCUSSION

Our study was based on the thirteenth edition of the Ethnologue [3], published in 1996, which lists more than 6700 languages spoken in 228 countries. We divided the countries in 12 groups (bins) according to area and then calculated the average diversity D of living languages in each bin. In Fig. 1 we plot our results for language diversity as a function of area. As one sees in this figure, the data points are well fitted by a power law:

$$D \sim A^z, \tag{1}$$

with $z = 0.41 \pm 0.03$, an exponent close to the largest values found in Ecology [7]. It should be emphasized that the power law shown in Fig. 1 extends over almost six decades, the only deviation occurring for countries with area smaller than 30 km^2. The density of living languages, $\rho_D = D/A$, thus scales as $\rho_D \sim A^{-0.59}$, implying that larger areas have proportionately less diversity of languages.

We have also studied how language diversity varies with population. In Fig. 2 we show the dependence of the average language diversity D as a function of the average population N within each area bin. In this case we find a power-law of the type:

$$D \sim N^\nu, \tag{2}$$

where from a best fit we obtain $\nu = 0.50 \pm 0.04$. It is interesting to note that similar scaling (with $\nu = 1/2$) between diversity and population has been found in computer simulations and experiments on fragmentation dynamics [6,10,12,13] as well as in insect populations [16,17]. Figure 2 also shows that on average a group of about 15,000 people is needed to maintain one single language alive. This might be of relevance *vis-à-vis* the potential danger of extinction of several languages [18] whose number of native speakers are presently well below this threshold.

From Eqs. (1) and (2) it follows that the (average) population grows with the (average) country area as a power law:

$$N \sim A^{z/\nu}. \tag{3}$$

From the values for z and ν above we thus obtain that $N \sim A^{0.82}$. (This result could also have been obtained directly from the data on country areas and populations without referring to the language distribution). In other words, the (average) population density on Earth, $\rho_N = N/A$ is not constant but rather decreases with area as $\rho_N \sim A^{-0.18}$. Thus countries with large areas are proportionately less populated than smaller ones, as is widely known. Moreover, from Eq. (3) and the fact that $A \sim L^2$, where L is a linear length scale, it then follows that

$$N \sim L^d, \tag{4}$$

where $d = 2z/\nu = 1.64$, thus indicating that the human population is distributed over the surface of the earth on a fractal set of dimension $d = 1.64$. Note also that from Eq. (1) and the fact that $A \sim L^2$ it follows that language diversity scales with linear size as

$$D \sim L^\delta, \tag{5}$$

where $\delta = 2z = 0.82$, meaning that living languages are distributed on a set of dimension close to unity.

Another interesting pattern concerns the distribution of language diversity among the various countries. We show in Fig. 3 the cumulative diversity distribution, $\mathcal{N}(>D)$, corresponding to the number of countries with a language diversity greater than D. We see from this figure that $\mathcal{N}(>D)$ displays composite power-laws: $\mathcal{N}(>D) \sim D^{-B}$, with $B = 0.6$ for $6 < D < 60$ and $B = 1.1$ for $60 < D < 700$ (each power-law in this case extends over one decade or more.) Now, why should the diversity distribution have two scaling regimes with $\mathcal{N}(>D)$ decreasing faster for larger D? A possible answer is that it is difficult in general to preserve the unity of large countries with great language (and hence ethnic) diversity, since they will tend to break up into smaller ones. This process could thus account, at least in part, for the fact that the cumulative diversity distribution $\mathcal{N}(>D)$ crosses over to a faster decay as D increases.

Scaling relations (1) and (2) above were obtained averaging the language diversity and the population over countries of similar area. To obtain a better estimate of the exponents z and ν one should ideally count the language diversity and total population contained in concentric regions, say, circles, of increasing size, as is costumary in statistical physics. Unfortunately, this procedure would be quite cumbersome here, if possible at all, and so we had to resort to the data reported in the Ethnologue [3] for individual countries. We believe, however, that the persistence of scaling behavior over several decades in Figs. 1 and 2 is an indication that our estimates are statistically reliable.

We have also studied the language size distribution—a quantity that does not directly depend on geopolitical boundaries. In Fig. 4 we show the cumulative size distribution, $n(>N)$, corresponding to the number of languages with a population greater than N. We see from this figure that $n(>N)$ displays composite power-laws:

$$n(>N) \sim N^{-\tau}, \tag{6}$$

with $\tau = 0.5$ for $5 \times 10^4 < N < 6 \times 10^6$, and $\tau = 1.0$ for $2 \times 10^7 < N < 1 \times 10^9$. Note that each of these power laws is valid for about two decades.

The fact that the exponents τ for the largest and intermediate-sized languages differ might be seen as an evidence that these two classes of languages possess distinct growth dynamics, leading to different patterns in the occupation of space. To see this, we first introduce the fractal dimension $\mathcal{D}$ defined by [9]:

$$n(>L) \sim L^{-\mathcal{D}}, \tag{7}$$

where $n(>L)$ is the number of languages that occupy a region of linear size greater than L. From Eqs. (4), (6), and (7) one then immediately finds :

$$\mathcal{D} = d\tau. \tag{8}$$

Thus, the languages with largest populations, for which $\tau = 1$, may be regarded as 'space filling' in the sense that $\mathcal{D} = d$, i.e., they are distributed on a subset of dimension equal to

the dimension of the set on which the entire population is distributed. On the other hand, languages with smaller populations ($\tau = 0.5$) are more 'tenuously' distributed on the surface of Earth, since they occupy a subset of dimension ($\mathcal{D} = d/2 = 0.82$) considerably less than d. Note also that in this case $\mathcal{D} = \delta$ [see Eq. (5)], thus showing that the dominant contribution to language diversity comes, as expected, from languages with small to intermediate-sized populations. The results above conform with the obvious fact that languages with large populations tend to be more widely spread, whereas languages with smaller populations are in general restricted to small areas (most of the languages in the region with $\tau = 0.5$ in Fig. 4 are indeed confined within a single country). It is also interesting to notice that the exponent $\tau = 1$ for languages with large populations is similar to what is usually found in classical critical phenomena [19]. Of course, a more detailed model for population dynamics is required if one wishes to explain, in a more quantitative fashion, the interesting features revealed by the present analysis. We are currently working on this direction.

III. CONCLUSIONS

We have presented a quantitative analysis of the diversity of human languages. We have studied how language diversity increases with area and population and found scaling relations in both cases. The language size distribution was also analyzed and shown to display two distinct power laws: (i) one with the exponent $\tau = 1$ for the top 50 languages by population and (ii) another one with $\tau = 0.5$ for languages with population between fifty thousand and six million. The corresponding fractal dimension $\mathcal{D}$ for these two classes of languages was obtained, and it was found that the largest languages are 'space filling' ($\mathcal{D} = d = 1.64$) with respect to the set available for the entire population, whereas intermediate-sized languages are more thinly distributed ($\mathcal{D} = d/2 = 0.82$) but give the main contribution to language diversity.

This work was partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico and Financiadora de Estudos e Projetos (Brazilian Agencies).

REFERENCES

[1] S. Blackmore, Science **274** (1996) 63.

[2] R. Doyle, Sci. Am. **279** (1998) 19.

[3] http://www.sil.org/ethnologue.

[4] D. W. Thompson, On Growth and Form, Cambridge University Press, Cambridge, 1971.

[5] S. Wolfram, Theory and Applications of Cellular Automata, World Scientific, Singapore, 1986.

[6] K. R. Coutinho, M. D. Coutinho-Filho, M. A. F. Gomes, A. N. Nemirovsky, Phys. Rev. Lett. **72** (1994) 3745.

[7] I. Hanski and M. Gyllenberg, *Science* **275** (1997) 397.

[8] D. M. Raup, S. J. Gould, T. J. Schopf, D. S. Simberloff, J. Geology **81** (1973) 525.

[9] B. B. Mandelbrot, The Fractal Geometry of Nature, Freeman, New York, 1983.

[10] K. Coutinho, S. K. Adhikari, M. A. F. Gomes, J. Appl. Phys. **74** (1993) 7577.

[11] C. Lewenkopf, J. Dreute, A. A.-Magd, J. Aichelin, W. Heinrich, J. Hüfner, G. Rusch, B. Wiegel, Phys. Rev. C **44** (1991) 1065.

[12] K. Coutinho, M. A. F. Gomes, S. K. Adhikari, Europhys. Lett. **18** (1992) 119.

[13] V. P. Brito, M. A. F. Gomes, F. A. O. Souza, S. K. Adhikari, Physica A **259** (1998) 227.

[14] M. L. Rosenzweig, Species Diversity in Space and Time, Cambridge Univ. Press, Cambridge, 1995.

[15] J. D. Pelletier, Phys. Rev. Lett. **82** (1999) 1983.

[16] E. Siemann, D. Tilman, J. Haarstad, Nature **380** (1996) 704.

[17] M. A. F. Gomes, G. L. Vasconcelos, S. K. Adhikari, I. J. Tsang, I. R. Tsang, Proceedings of the XV SITGES Euroconference–Statistical Mechanics of Biocomplexity, Barcelona, June 1998, pp. 33-35.

[18] J. E. Grimes, Oceanic Linguistics **34** (1995) 1.

[19] M. E. Fischer, Proc. Int. Sch. Phys. Enrico Fermi, Course LI, Critical Phenomena, M. S. Green (Ed.), Academic Press, New York, 1971.

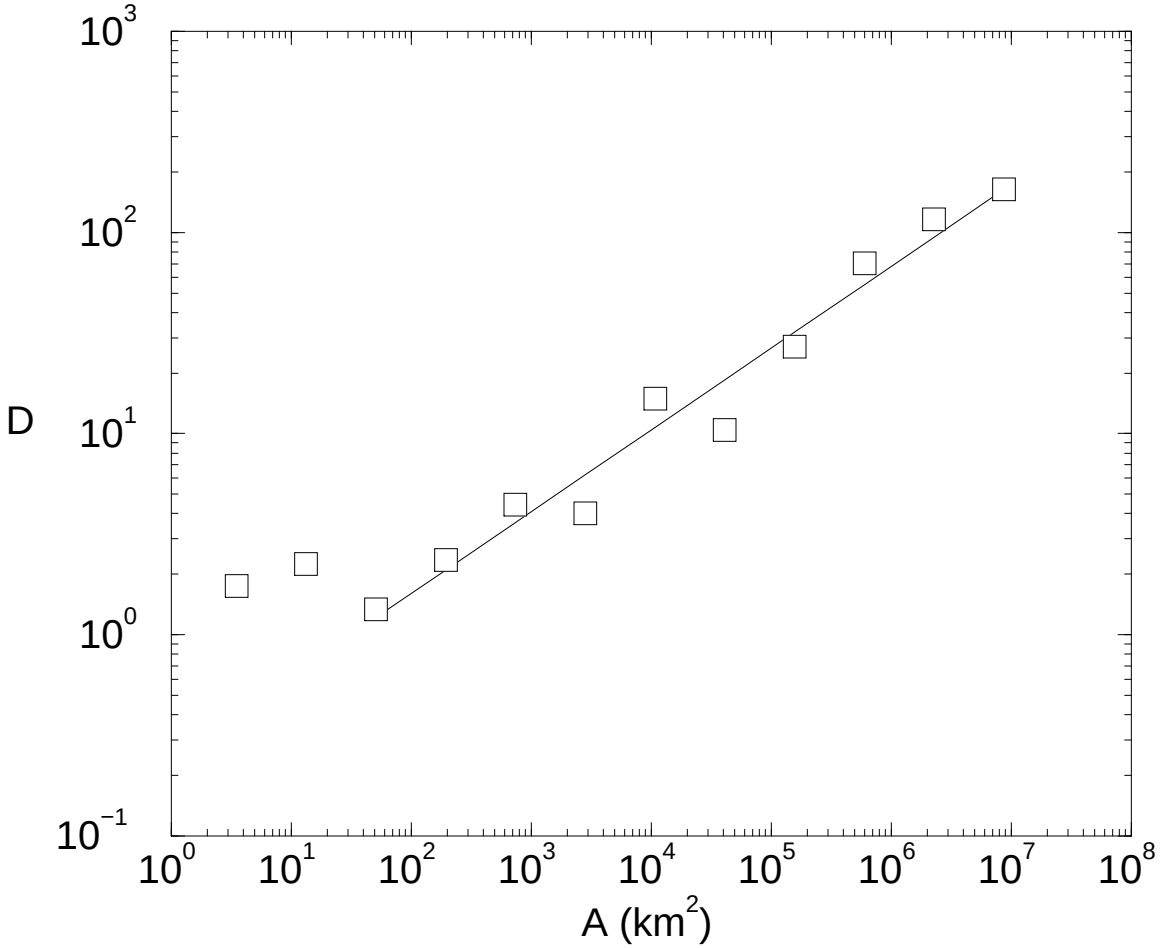
FIGURES

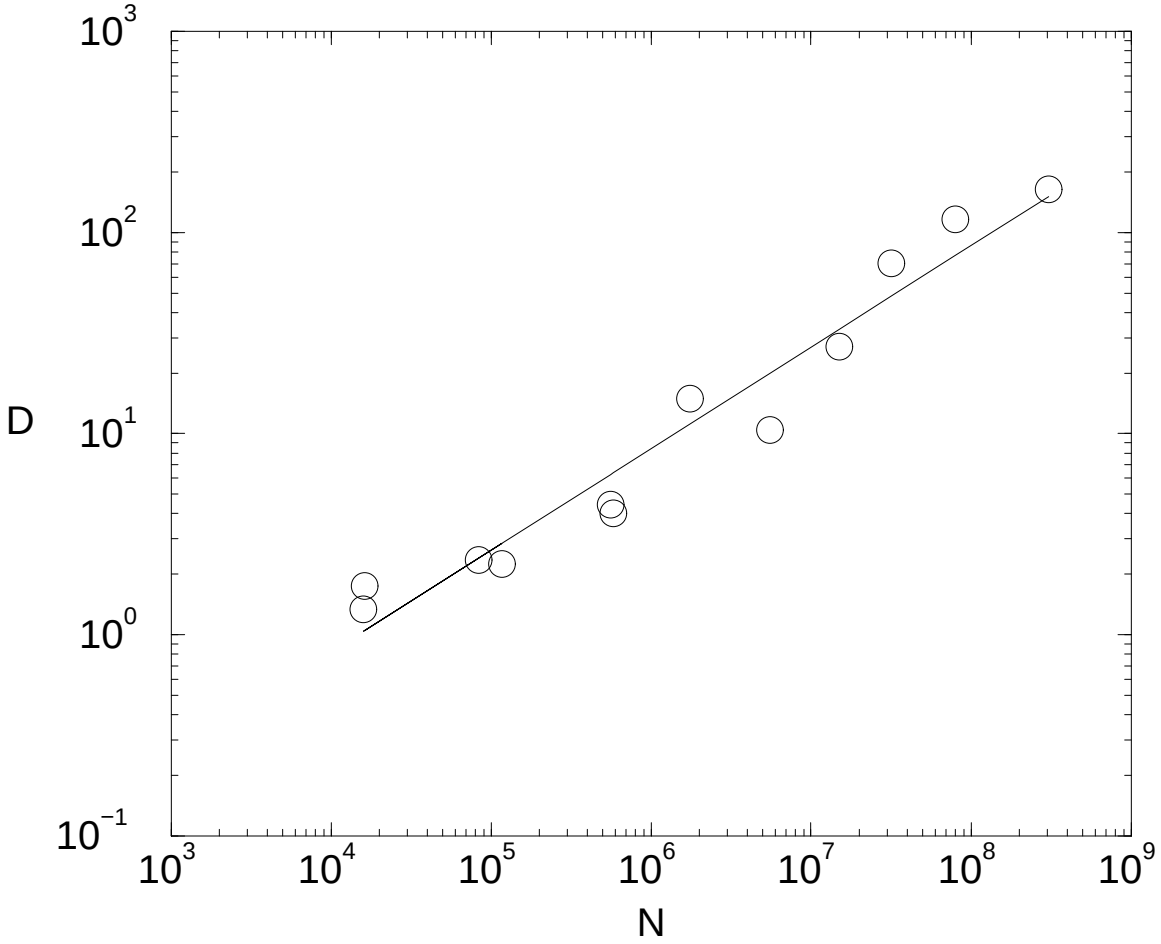
FIG. 1. Average diversity of languages as a function of area. The straight line is a best fit whose slope gives the exponent $z = 0.41 \pm 0.03$.

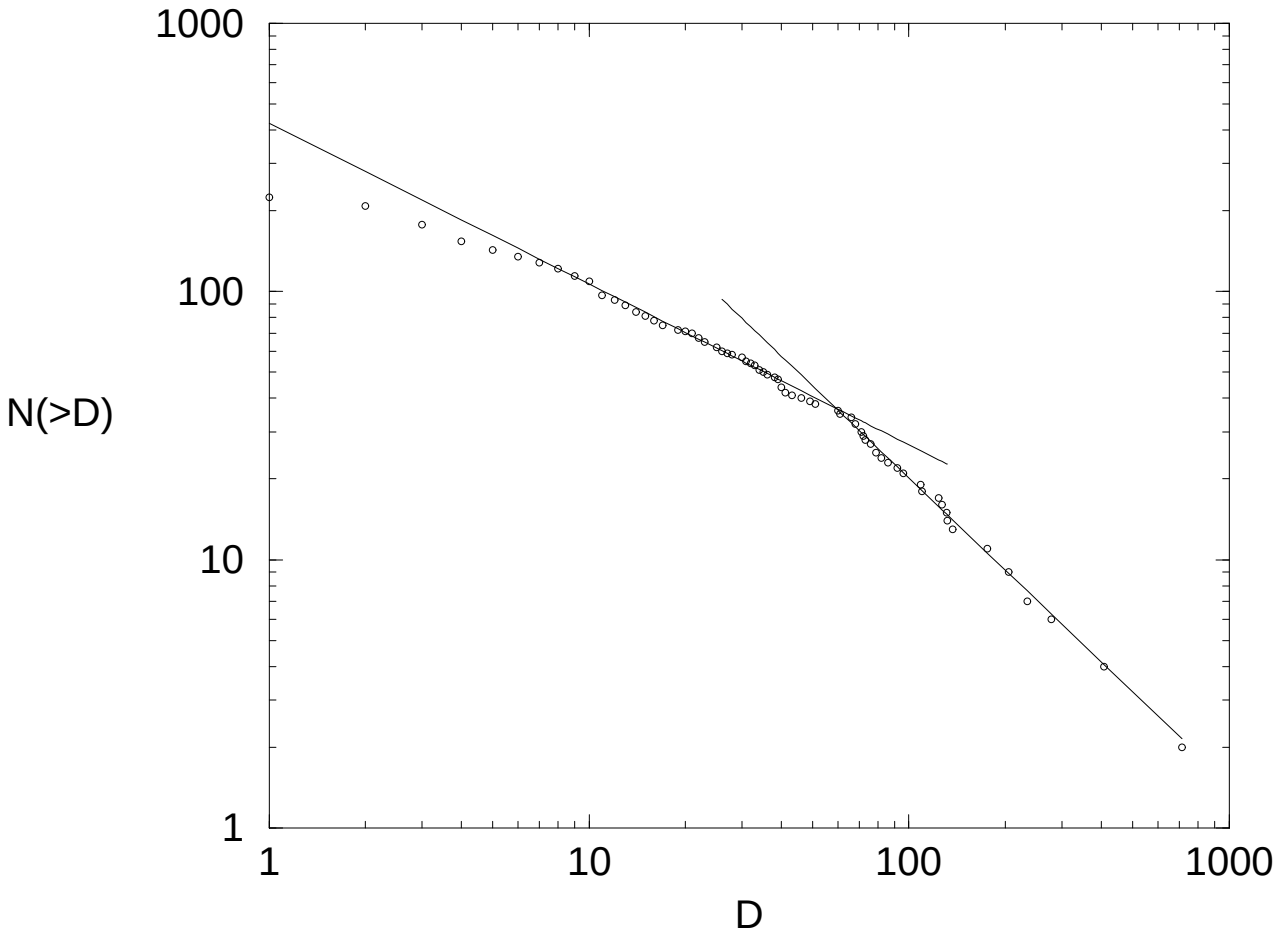
FIG. 2. Average diversity of languages as a function of average population within a bin area. The solid line is a best fit with slope $\nu = 0.50 \pm 0.04$.

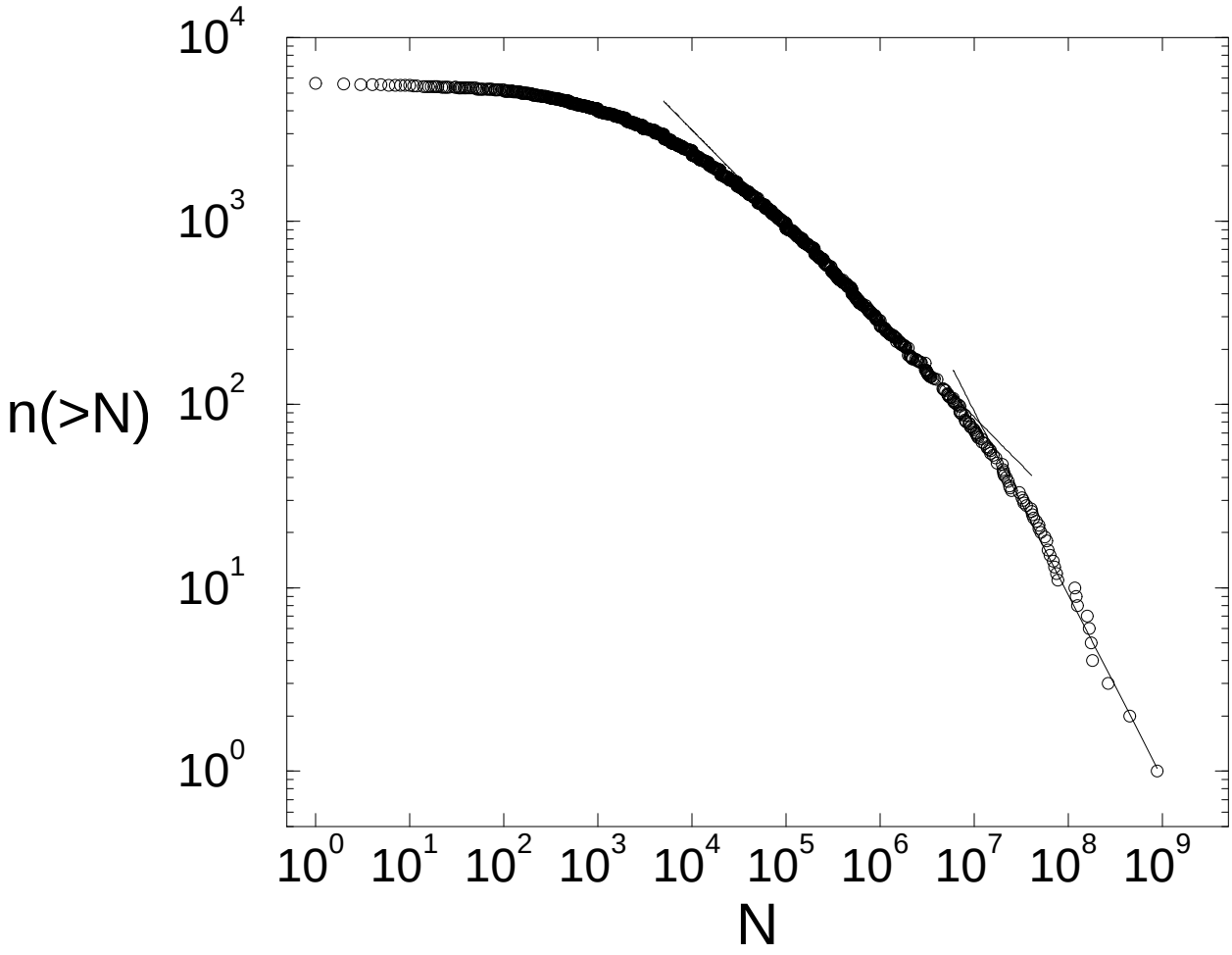
FIG. 3. Number of countries with a language diversity greater than D as a function of D. The straight lines give the exponents $B = 0.6$ for $6 < D < 60$ and $B = 1.1$ for $60 < D < 700$.

FIG. 4. Number of languages with population greater than N as a function of N. The straight lines give the exponents $\tau = 0.5$ for $5 \times 10^4 < N < 6 \times 10^6$ and $\tau = 1.0$ for $2 \times 10^7 < N < 1 \times 10^9$.


D
A (km^2)
10^3
10^2
10^1
10^0
10^{-1}
10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7 10^8


10^3
10^2
10^1
10^0
10^{-1}
D
10^3
10^4
10^5
10^6
10^7
10^8
10^9
N


1000
100
10
1
N(>D)
1
10
100
1000
D


n(>N)
10^4
10^3
10^2
10^1
10^0
10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7 10^8 10^9
N