International Journal of Modern Physics C

DEMOGRAPHIC GROWTH AND THE DISTRIBUTION OF LANGUAGE SIZES

DAMIÁN H. ZANETTE

Consejo Nacional de Investigaciones Científicas y Técnicas
Centro Atómico Bariloche and Instituto Balseiro
8400 San Carlos de Bariloche, Río Negro, Argentina.
zanette@cab.cnea.gov.ar

Received Day Month Year
Revised Day Month Year

It is argued that the present log-normal distribution of language sizes is, to a large extent, a consequence of demographic dynamics within the population of speakers of each language. A two-parameter stochastic multiplicative process is proposed as a model for the population dynamics of individual languages, and applied over a period spanning the last ten centuries. The model disregards language birth and death. A straightforward fitting of the two parameters, which statistically characterize the population growth rate, predicts a distribution of language sizes in excellent agreement with empirical data. Numerical simulations, and the study of the size distribution within language families, validate the assumptions at the basis of the model.

Keywords: Language evolution; population dynamics; multiplicative stochastic processes.

PACS Nos.: 87.23.Ge, 89.65.Cd, 02.50.Ey

1. Introduction

Statistical aspects of the evolution of languages have attracted, in the last few years, a great deal of attention among physicists and mathematicians.[1,2,3,4,5,6,7,8,9,10,11] One of the better established quantitative empirical facts about extant languages is their size distribution, namely, the frequency of languages with a given number of speakers. Naturally, explaining the origin of this distribution is one the aimed goals of mathematical modeling in this field.

Recent work has built up on variations of two basic models of language evolution, Schulze's model[5] and Viviane's model,[7] both of them mostly focused on the effects of mutation of linguistic features which give rise to new languages, and including the possibility of language extinction. These models, however, disregard the fact that, over periods which are short as compared with the typical time scales of language evolution, the speakers of a given language can substantially vary in number just by the effect of population dynamics. For instance, in the last five centuries –a period which includes the culturally devastating European invasion of the rest of the globe– perhaps 50 % of the world's languages went extinct (among them, two thirds of the

2,000 preexisting native American languages[12]) or changed drastically. In the same period, however, the world's population grew by a factor of twelve or more.[13]

Demographic effects have been very recently incorporated to a model of language evolution by Tuncay,[11,14] in the form of a stochastic multiplicative model for population growth (see also Ref. 15). With suitable tuning of its several parameters, Tuncay's model is able to reasonably reproduce the observed distribution of language sizes, as a result of numerical simulations.[11] The complexity of the model –which, in its full form,[14] includes population growth along with language inheritance, branching, assimilation, and extinction– makes it however difficult to identify the specifc mechanism that shapes the distribution of sizes.

In this paper, I show that the empirical distribution of language sizes can be accurately explained taking into account just the effect of demographic processes. This possibility was already pointed out, for the specific case of the languages of New Guinea, by Novotny and Drozd.[16] I propose a two-parameter stochastic model where the populations speaking different languages evolve independently of each other. During the evolution –which, in the realization of the present model, is assumed to span 1,000 years– language creation, assimilation, and extinction are disregarded. This assumption does not discard mutations inside a given language, which may lead to its internal evolution, but each language preserves its identity as a cultural and demographic unit along the whole period. The model is analytically tractable, and its two parameters can be fitted *a priori* from empirical data. Numerical simulations confirm the prediction that the distribution is essentially independent of details in the initial condition –the distribution of sizes ten centuries ago– so that, in a sense, the present distribution is the unavoidable consequence of just demographic growth. The model is further validated by showing that the distribution of sizes of languages belonging to a given family has the same shape as the overall distribution. These results strongly suggest that population dynamics is a necessary ingredient in models of linguistic evolution.

2. Present distribution of language sizes

It is a well established empirical fact that the frequency of languages with a given number of speakers p is satisfactorily approximated by a log-normal distribution.[17,5] Accordingly, the distribution of language sizes as a function of the log-size, $q = \ln p$, is approximately given by a Gaussian,

$$Q(q) = (2\pi\theta^2)^{-1/2} \exp\left[-(q-\bar{q})^2/2\theta^2\right], \tag{1}$$

where $\bar{q}$ and θ are, respectively, the mean value and the mean square dispersion of q. The quantity $Q(q)dq$ gives the fraction of languages with log-sizes in $(q, q + dq)$. Significant departures from the log-normal distribution are limited to small language sizes, up to the order of a few tenths speakers.[17]

Ethnologue statistical summaries,[18] whose data correspond to collections done mostly during the 1990s, list the number of languages with sizes within decade bins

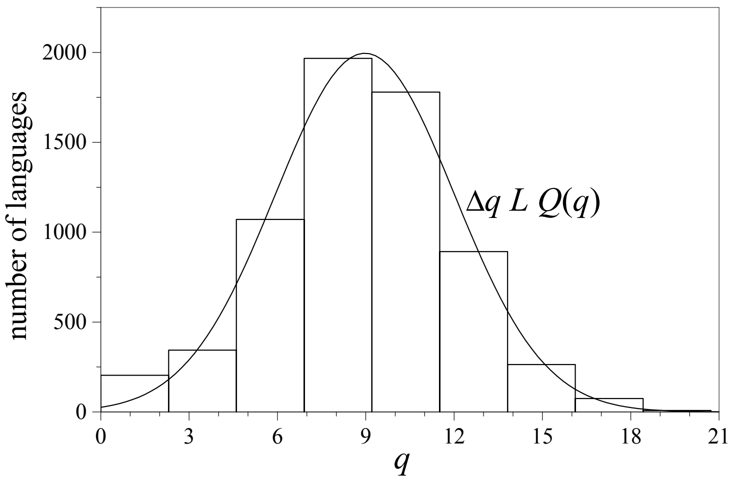

Fig. 1. The histogram shows the number of languages whose sizes p have logarithms $q = \ln p$ in the corresponding bins, according to the Ethnologue statistical summaries.[18] The curve is the Gaussian distribution $Q(q)$ of Eq. (1) with the parameters of Eq. (4). For comparison with the histogram, $Q(q)$ is multiplied by the total number of languages, $L = 6,604$, and by the bin width $\Delta q = \ln 10$. Note the deviation from normality in the leftmost column of the histogram.[17]

(1 to 9 speakers, 10 to 99 speakers, 100 to 999 speakers, and so on), and give the number of speakers within each bin. The total number of languages in the list is $L = 6,604$, accounting for an overall population a little above 5.7×10^9 speakers. The sizes of 308 languages of the database are unknown. In terms of the distribution $Q(q)$, the number of languages in the bin between 10^k and 10^{k+1} speakers is

$$L_k = L \int_{k \ln 10}^{(k+1)\ln 10} Q(q)\ dq, \tag{2}$$

while the total population speaking the languages in the same bin is

$$p_k = L \int_{k \ln 10}^{(k+1)\ln 10} \exp(q)\ Q(q)\ dq. \tag{3}$$

The values of L_k and p_k provided by the Ethnologue statistical summaries can thus be used to estimate the parameters $\bar{q}$ and θ in the distribution $Q(q)$ of Eq. (1). Least-square fitting yields

$$\bar{q} = 8.97 \pm 0.15, \qquad \theta = 3.04 \pm 0.09. \tag{4}$$

Figure 1 shows, as a histogram over the variable q, Ethnologue's data for L_k. The curve is a plot of the function $LQ(q)$ with the parameters of Eq. (4), whose integral over the histogram bins approximates the empirical values of L_k. For easier comparison with the histogram, $LQ(q)$ is further multiplied by the bin width $\Delta q =$

$\ln 10$. The aim in the following is to provide a model which explains the fitted distribution $Q(q)$.

3. Demographic evolution of language sizes

The log-normal shape of the distribution of language sizes suggests that a multiplicative stochastic process is at work in the evolution of the number of speakers of each language. This is in turn consistent with the hypothesis that, over sufficiently long time scales, the population speaking a given language evolves autonomously, driven just by demographic processes.[16]

Let $p_t^{(j)}$ be the number of speakers of language j at time t, and assume that at time $t+1$ –one year later, say– the population has changed to[11]

$$p_{t+1}^{(j)} = \alpha_t^{(j)} p_t^{(j)}, \tag{5}$$

where the growth rate $\alpha_t^{(j)}$ is a positive stochastic variable drawn from some specified distribution. As shown below, the mean value and the mean square dispersion over this distribution can be estimated from empirical data. In terms of the initial population $p_0^{(j)}$, the number of speakers at time t is

$$p_t^{(j)} = p_0^{(j)} \prod_{s=0}^{t-1} \alpha_s^{(j)}. \tag{6}$$

I suppose now that, during the whole t-step process, the distribution of the growth rate $\alpha_t^{(j)}$ is (i) the same for all languages, and (ii) does not vary with time. Moreover, (iii) no language is created or becomes extinct. Admittedly, these are rather bold assumptions for the world's history during the last 1,000 years. However, in view of the lack of reliable data over such period, they are at least justified by the sake of simplicity.

I identify the evolution of the world's languages as $L = 6,604$ realizations of the multiplicative stochastic process (5). By virtue of assumption (i), all the realizations are statistically equivalent. In this interpretation, the present distribution of log-sizes, given by Eqs. (1) and (4), is the probability distribution for the variables $p_t^{(i)}$ obtained from those realizations. Thus, my aim is to quantitatively relate the distribution $Q(q)$ to the outcome of the stochastic process.

The total population P_t at time t is

$$P_t = \sum_{j=1}^{L} p_t^{(j)} = \sum_{j=1}^{L} p_0^{(j)} \prod_{s=0}^{t-1} \alpha_s^{(j)}. \tag{7}$$

Averaging this expression over realizations of the stochastic variable $\alpha_t^{(j)}$, and assuming that the growth rate is not self-correlated in time, we find $\langle P_t \rangle = \langle \alpha \rangle^t P_0$, where $\langle \alpha \rangle$ is the mean growth rate and P_0 is the initial total population. In order to apply this analysis to the world's population in the last ten centuries, let us take $P_0 = 3.1 \times 10^8$, which is the estimated population by the year 1000.[13] Ascribing the total population accounted for by Ethnologue's data to the year 2000, and associating this number with the population averaged over realizations of the growth rate, we have $\langle P_t \rangle = 5.7 \times 10^9$ and $t = 10^3$. This makes it possible to evaluate the mean growth rate per year as

$$\langle \alpha \rangle = (\langle P_t \rangle / P_0)^{1/t} \approx 1.0029. \tag{8}$$

To evaluate the dispersion of the growth rate, it is useful to introduce its relative deviation with respect to the average, $\delta_t^{(j)}$, as

$$\alpha_t^{(j)} = \langle \alpha \rangle \left[1 + \delta_t^{(j)} \right]. \tag{9}$$

The average value and the mean square dispersion of the deviation $\delta_t^{(j)}$ are, respectively,

$$\langle \delta_t^{(j)} \rangle = 0, \qquad \sigma_\delta \equiv \langle \delta_t^{(j)2} \rangle^{1/2} = \sigma_\alpha / \langle \alpha \rangle, \tag{10}$$

where σ_α is the mean square dispersion of the growth rate. Assuming that $\delta_t^{(j)}$ is always sufficiently small to approximate $\ln(1 + \delta_t^{(j)}) \approx \delta_t^{(j)} - \delta_t^{(j)2}/2$, Eq. (6) can be rewritten for the log-size $q_t^{(j)} = \ln p_t^{(j)}$ as

$$q_t^{(j)} = q_0^{(j)} + t \ln \langle \alpha \rangle + \sum_{s=0}^{t-1} \left[\delta_s^{(j)} - \delta_s^{(j)2}/2 \right]. \tag{11}$$

This equation shows explicitly that, besides the deterministic growth given by the term $t \ln\langle \alpha \rangle$, the evolution of the logarithm of the population speaking a given language is driven by an additive stochastic process. Thus, by virtue of the central limit theorem,[19] the distribution $Q(q)$ must converge to a Gaussian like in Eq. (1), starting from any distribution of initial log-sizes $q_0^{(j)}$. For the time being, however, the question remains whether the times relevant to the process are enough to allow for the development of the Gaussian shape and, in particular, to suppress any effect ascribable to a specific initial distribution.

Unfortunately, the initial sizes $p_0^{(j)}$ –the number of speakers of each language 1,000 year ago– are not known for most languages. However, their effect on the present size distribution can be readily assessed. Averaging Eq. (11) over realizations of the stochastic process and over the distribution of initial log-sizes –and always assuming that the deviations $\delta_t^{(j)}$ are small– yields, for the mean square dispersion of $q_t^{(j)}$,

$$\sigma_q^2 = \sigma_0^2 + \sigma_\delta^2 t, \tag{12}$$

where σ_0 is the mean square dispersion of the initial log-sizes. The empirical estimation for σ_q is the value of θ given in Eq. (4). In turn, an upper bound can be given for σ_0, as the maximal mean square dispersion in the log-size distribution of $L = 6,604$ languages with a total population of $P_0 = 3.1 \times 10^8$ speakers, and at least one speaker per language. This maximal dispersion is obtained with $L-1$ languages with exactly one speaker, and just one language with the remaining $P_0 - L + 1$ speakers. Clearly, this is an unlikely distribution for the languages 1,000 years ago, but represents the worst-case instance, with the largest contribution of the initial condition to the present dispersion of log-sizes. In this extreme situation, the estimation for the initial mean square dispersion is $\sigma_0^2 \approx L^{-1} \ln^2 P_0 \approx 0.058$. Meanwhile, according to Eq. (4), $\sigma_q^2 \approx 9.2$. In the right-hand side of Eq. (12), therefore, the largest contribution by far is given by the second term, and that of the initial distribution is essentially negligible. This makes it possible to calculate the mean square dispersion of the deviations $\delta_t^{(j)}$ as

$$\sigma_\delta \approx t^{-1/2}\theta \approx 0.096 \tag{13}$$

for $t = 10^3$, thus completing the statistical characterization of the growth rate per year, $\alpha_t^{(j)}$. Note that this value of σ_δ is in agreement with the assumption of small relative deviations in the growth rate.

Summarizing the results of this section, I have argued that the present log-normal distribution of language sizes can be seen as the natural consequence of population dynamics driven by a stochastic multiplicative process, Eq. (5), where the evolution of each language is interpreted as a realization of the process. Using data on the total population growth during the last 1,000 years –a period over which I neglect language birth and death– and fitting only one parameter (σ_δ) from the distribution itself, I was able to statistically characterize the growth rate per year which explains the present distribution, giving its mean value and mean square dispersion. Also, I have advanced that the dispersion of language sizes ten centuries ago has essentially no effect on its present value. It is now useful to validate these conclusions with numerical realizations of the model, and with applications within language families.

4. Validation of the model

4.1. *Numerical results*

In this section, I present results for series of $L = 6,604$ numerical realizations of the multiplicative stochastic process (5). The mean value and the mean square dispersion of the growth rate $\alpha_t^{(i)}$ are fixed according to the values estimated in Section 3, Eqs. (8) and (13). In order to speed up the computation, individual values of $\alpha_t^{(i)}$ are drawn from a square distribution centered at $\langle\alpha\rangle$, with a width which insures the correct mean square dispersion. In agreement with my main assumptions, I avoid the possibility that languages die out by replacing the absorbing boundary at $p = 1$, below which a language should become extinct, by a reflecting boundary.

In view of the discussion in the previous section, the convergence of the distribution of log-sizes to a Gaussian is guaranteed by the central limit theorem. The emphasis in the simulations is thus put on the possible effects of the distribution of initial sizes $p_0^{(i)}$. Figure 2 shows, as normalized histograms, numerical results for single series of L realizations of the stochastic process after $t = 10^3$ steps, from four different initial conditions. In (a), all the languages have exactly the same initial size p_0. In (b), the initial sizes are uniformly distributed between $p = 0$ and $p_{\max}$. In (c), the distribution of initial sizes is also uniform, but spans the interval $(p^*, 2p^*)$. Finally, in (d) the initial distribution is more heterogeneous, with half the languages having size $p^\dagger$ and the other half having size $10p^\dagger$. The parameters p_0, $p_{\max}$, p^*, and $p^\dagger$, which characterize these initial distributions, are fixed by the condition that the total population is $P_0 = 3.1 \times 10^8$. The curve in all plots is the Gaussian of Eq. (1) with the parameters of Eq. (4). The agreement in cases (a) to (c) is excellent. Only in case (d), where the initial distribution is specially heterogeneous –and, certainly, not a likely representation of the distribution of language sizes ten centuries ago– the deviations are larger, though the agreement is still very reasonable.

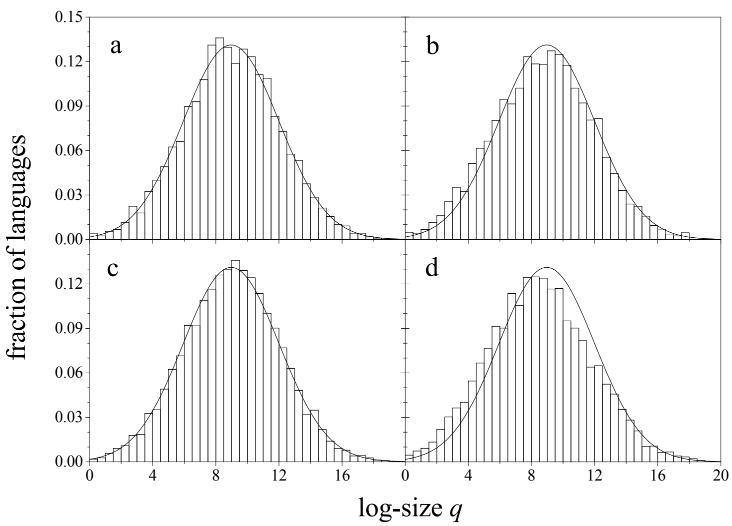

Fig. 2. Normalized histograms for $L = 6,604$ language sizes after 10^3 steps of the stochastic process (5), starting from the four initial conditions described in the text. Curves stand for the expected Gaussian distribution, Eq. (1), with the parameters of Eq. (4).

The difference between the numerical results of case (d) and the expected Gaussian function resides not only in the width of the distribution but also in its mean value. To a much lesser extent, this discrepancy is also visible in case (b). This shift between the distribution peaks can be understood in terms of the average of Eq. (11) over both the realizations of the growth rate $\alpha_t^{(i)}$ and the initial log-sizes $q_0^{(i)}$,

$$\langle q_t^{(i)} \rangle = \langle q_0^{(i)} \rangle + t(\ln\langle \alpha \rangle - \sigma_\delta^2/2). \tag{14}$$

Besides the contribution of the multiplicative stochastic process, given by the term proportional to the time t, the mean value $\langle q_t^{(i)} \rangle$ in Eq. (14) depends on the average initial log-size $\langle q_0^{(i)} \rangle$. In spite of the fact that the total initial population and the number of languages are the same for all simulations –which always gives the same average size per language– the average log-size depends on the specific initial distribution. Thus, the final mean values for different initial conditions are generally shifted with respect to each other.

As a consistency test for the assumption that languages do not die out along the evolution period considered here, I have also run simulations taking into account the absorbing boundary at $p = 1$. Namely, when the size of a language decreases below one speaker during its evolution, it is considered to become extinct and that particular realization of the stochastic process is interrupted. Among the four initial conditions considered above, those who undergo larger extinctions are, not unexpectedly, (b) and (d) –as they produce the largest shifts to the left in the log-size distributions. In both cases, however, the fraction of extinct languages is around 1 %, which validates the above assumption quite satisfactorily.

4.2. *Size distribution within language families*

A crucial consequence of the hypotheses on which the present model is based –in particular, the mutual independence of the size evolution of different languages– is that its predictions should hold not only for the ensemble of all the world's languages, but also for any sub-ensemble to which the homogeneity assumptions (i) and (ii) reasonably apply. In other words, the final log-normal shape of the size distribution should also result from the evolution of, say, the languages of a given geographical region, or belonging to a given language family. This can be readily assessed from empirical data on the number of speakers of individual languages and, in fact, has already been pointed out for a set of some 1,000 New Guinean languages.[16]

Here, I analyze the size distribution of languages belonging to each one of the four largest families, according to Ethnologue's classification.[18] Population data for individual languages were obtained from Ethnologue's online databases. Figure 3 shows histograms of log-sizes for those families. To ease the comparison, the column width and the horizontal scale are the same as in Fig. 1. Curves stand for least-square fittings with Gaussian functions as in Eq. (1). The resulting parameters are $\bar{q} = 10.3$ and $\theta = 2.3$ for Niger-Congo; $\bar{q} = 8.2$ and $\theta = 2.4$ for Austronesian; $\bar{q} = 7.2$ and $\theta = 2.0$ for Trans-New Guinea; and $\bar{q} = 9.8$ and $\theta = 3.5$ for Indo-European. The quality of the agreement between the data and the Gaussian fitting is clearly comparable to that of the whole language ensemble in Fig. 1.

Note the interesting fact that the mean square dispersion θ –which, according to the present model, results from the dispersion in the population growth rate– is sensibly larger for the Indo-European family than for the other three. Surely, this is a direct consequence of the highly diverse fate of European languages in the last few centuries. In any case, the four mean square dispersions are not far from the overall value given in Eq. (4).

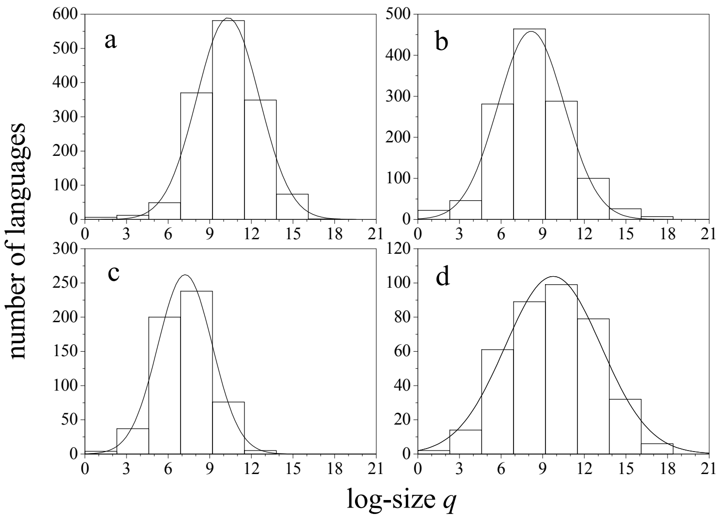

Fig. 3. Histograms of the number of languages as a function of the language log-size for four language families: (a) Niger-Congo (1,495 languages), (b) Austronesian (1,246 languages), (c) Trans-New Guinea (561 languages), and (d) Indo-European (430 languages). Curves are Gaussian least-square fittings.

5. Conclusion

In this paper, I have argued that the present log-normal distribution of language sizes is essentially a consequence of demographic dynamics in the population of speakers of each language. In fact, an isolated population can largely vary in number within time scales which are short as compared with those involved in substantial language evolution. To support this suggestion, I have proposed a stochastic multiplicative process for the population dynamics of individual languages, where language birth and death are disregarded. Within some bold assumptions on the geographical and temporal homogeneity of the process, the model is completely specified by two parameters, which give the average growth rate of the population and its mean square dispersion. The average growth rate is completely defined by the initial and the final world population. I have chosen to apply the model on the period spanning the last 1,000 years, for which reliable data on the world's population growth are available. The mean square dispersion of the growth rate is the only parameter which I fitted in an admittedly *ad-hoc* manner, using the present distribution of language sizes. It seems unlikely to find estimations for this parameter from independent historical sources, which would require to have reliable records on the population change year by year. Note that the dispersion in the growth rate is determined by a variety of factors, including fluctuations in birth and mortality frequencies and migration events.[15]

Once the two parameters are fitted, the model is able to produce, as the result of evolution along ten centuries, excellent predictions of the present distribution of language sizes. Numerical simulations show that the final distribution is largely independent on the initial condition. This emphasizes the point that, irrespectively of the long-range historical processes that may have determined the distribution of language sizes of 1,000 years ago –including language birth and death, branching, mutation, competition, assimilation, and/or replacement– population dynamics is by itself able to explain the present distribution. This conclusion had already been advanced for the case of New Guinean languages in Ref. 16. In fact, realizing that the same log-normal profile is found in the size distribution inside language families, is a further validation of the present model. In view of the present arguments, one can moreover safely assert that the distribution of language sizes was already a log-normal function, with different parameters, in year 1000.

It is clear that in the last few years –with the advent of a host of new mechanisms of globalization which endanger cultural diversity– many, or most, of the world's languages are threatened by the risk of extinction.[4,17,20] This risk is peculiarly acute for those languages whose number of speakers is below a few hundreds – including the range where the distribution of sizes differs from the log-normal profile (cf. Fig. 1). It should be a program of obviously urgent interest to study in detail what are the relevant processes at work in that range, even if they escape the domain of the statistical physicists' approaches.

Acknowledgments

I am grateful to Susanna C. Manrubia, who brought to my attention the significance of the distribution of language sizes to the problem of language evolution in 1996. Also, I acknowledge critical reading of the manuscript by A. A. Budini.

References

1. M. A. Nowak and D. C. Krakauer, *Proc. Nat. Acad. Sci. U. S. A.* **96**, 8028 (1999).
2. D. H. Zanette, *Adv. Compl. Sys.* **4**, 281 (2001).
3. R. Ferrer i Cancho and R. V. Solé, *Proc. Nat. Acad. Sci. U. S. A.* **100**, 788 (2003).
4. D. M. Abrams and S. H. Strogatz, *Nature* **424**, 900 (2003).
5. C. Schulze and D. Stauffer, *Int. J. Mod. Phys. C* **16**, 781 (2005).
6. V. Schwämmle, *Int. J. Mod. Phys. C* **16**, 1519 (2005).
7. V. M. de Oliveira, M.A.F. Gomes and I. R. Tsang, *Physica A* **361**, 257 (2006).
8. S. Wichmann, D. Stauffer, F. W. S. Lima, and C. Schulze, *Trans. Philol. Soc.* **105**, 126 (2007).
9. D. Stauffer, X. Castelló, V. M. Eguíluz and M. San Miguel, *Physica A* **374**, 835 (2007).
10. P. M. Castro de Oliveira, D. Stauffer, F. W. S. Lima, A. O. Sousa, C. Schulze, S. Moss de Oliveira, *Physica A* **376**, 609 (2007).
11. Ç. Tuncay, *Int. J. Mod. Phys. C* **18**, 1203 (2007).
12. L. Campbell and M. Mithun, eds., *The Languages of Native America: A Historical and Comparative Assessment* (University of Texas Press, Austin, 1979). See also http://www.nativeamericans.com/Natives.htm.
13. United Nations, Department of Economic and Social Affairs, Population Division, *The World at Six Billion*, http://www.un.org/esa/population/publications/sixbillion/.

14. Ç. Tuncay, arXiv:0705.1838.
15. D. Stauffer, C. Schulze, F. W. S. Lima, S. Wichmann, and S. Solomon, *Physica A* **371**, 719 (2007).
16. V. Novotny and P. Drozd, *Proc. Roy. Soc. London B* **267**, 947 (2000).
17. W. J. Sutherland, *Nature* **423**, 276 (2003).
18. R. G. Gordon, ed., *Ethnologue: Languages of the World*, 15th edition (SIL International, Dallas, 2005). Online version: http://www.ethnologue.com/.
19. N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1992).
20. M. Krauss, *Language* **68**, 4 (1992).