Universal Evolutionary Rate of Human Languages

P.M.C. de Oliveira[1], A.O. Sousa[2] and S. Wichmann[3]

1) Instituto de Física, Universidade Federal Fluminense, and INCT-SC, Brazil, pmco@if.uff.br

2) Departamento de Física, Universidade Federal do Rio Grande do Norte, Brazil, aosousa@dfte.ufrn.br

3) Max Planck Institute for Evolutionary Anthropology and Leiden University, Germany, wichmann@eva.mpg.de

abstract

We define the concept of spectrum for an evolutionary tree and its degeneracies, inspired on the energy spectrum of quantum mechanical systems. By plotting this spectrum according to a ranking procedure, one gets a decaying staircase with straight steps. The same kind of plot can be constructed with the pair distances between real spoken languages measured within a given family. In this case we find a bent, decaying curve where an elbow always appears. By comparing the theoretical staircase plot with the real bent curves, one can interpret the elbow as defining the time when the very first historical bifurcations occurred, for instance when a family bifurcates into genera. The distance between two languages is traditionally measured by linguists according to well defined procedures. We use data of such distances for a set of 89 language families. A careful analysis of their ranking plots and the corresponding elbows allowed us to estimate, for each family, the moments in the past when two events occurred: 1) the birthday of the family itself (F millennia ago); 2) the moment of its first bifurcations (G millennia ago). Then, by plotting G against F, one point per family, we find a straight line with regression slope $\Delta G/\Delta F = 1.021 \pm 0.047$. This unitary slope means that all families evolve (and bifurcate) according to the same, universal time rate. We interpret this as a result of the human speaking abilities, independent of the particular language structure of each spoken language or language family.

Figure 1 (left) shows an example of evolutionary tree, time runs up-down. Bottom black points correspond to alive individuals (languages), while red bullets are their past ancestors. For any pair of current languages, one can go back in time along each branch, in order to find its common ancestor. The length of this backward path until the common ancestor is the theoretical ultrametric distance (u-distance) between the corresponding pair of languages. In other words, the u-distance measures the common ancestor's age.

Figure 1 (right) shows the corresponding spectrum, the set of u-distances (or past bifurcation events) taken for all 15 pairs of current languages. Each level (horizontal segment) corresponds to one or more pairs of current languages. In this example, the uppermost level corresponds to 8 different pairs, i.e. the same common ancestor for all these 8 pairs. This level is said to be 8-fold degenerate. The second uppermost level is 4-fold degenerate.

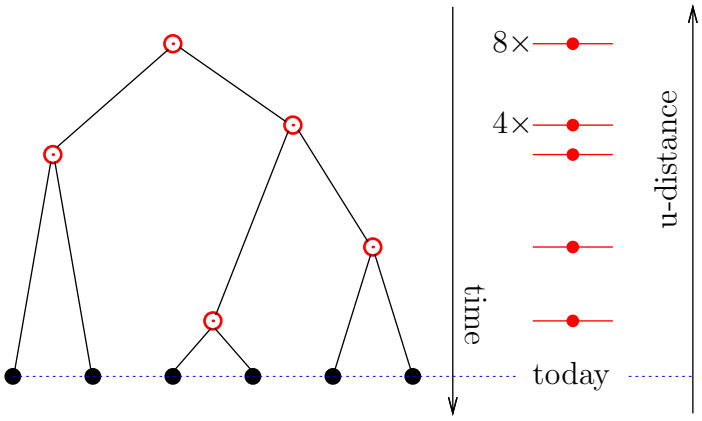

Figure 1: Example of an evolutionary tree and its spectrum.

Figure 2 shows the so-called ranking plot, where the u-distances are simply displayed in decreasing order.

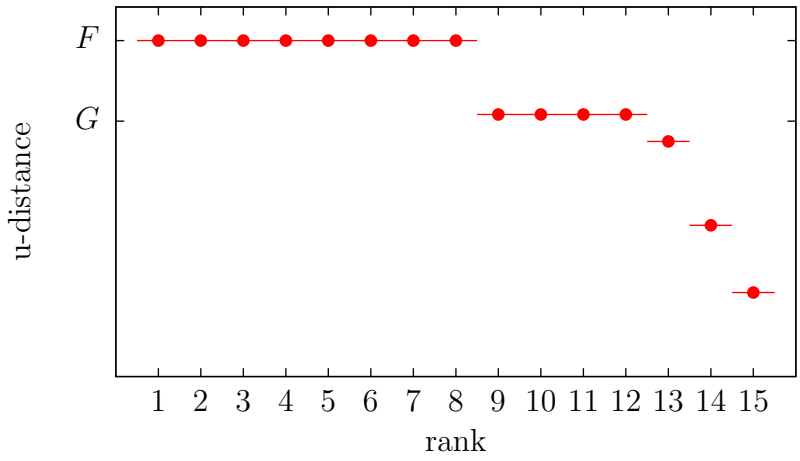

Figure 2: Ranking plot for the same 15 u-distances shown in Figure 1. We denote by F (family) and G (genera) the values corresponding to the uppermost level and the average of the next two, respectively. They correspond to the age of the whole family and when it bifurcates in genera.

This is only theory, since the knowledge of the past moments when ancestor languages bifurcate are not available, except for very few and recent particular cases. Thus, u-distances are not directly measurable. However, they are indirectly measured by comparing a set of well chosen words spoken today in one language with the corresponding words spoken in the other language. This procedure is well stablished among linguists, since the pioneering work of Swadesh [1]. After each bifurcation, both branches are supposed to evolve independently from each other, therefore the older is the common ancestor of two currently spoken languages the larger will be the distance obtained from such a comparison. These indirectly measured distances, however, suffer from fluctuation effects due to random drifts during the past evolution. Being and approximation to the theoretical u-distances, they do not follow a precise staircase as exemplified in Figure 2 with its pre-

cise horizontal plateaux. Instead, the ranking plot for a real language family is exemplified in Figure 3, where the distances are measured according to the traditional method introduced by Levenshtein [2]. Distances are normalized by dividing all of them by a standard value obtained from other languages known a priori to be independent of each other, belonging to different families [3, 4]. Therefore, the largest conceivable distance within the same family is 1. The values at the very left in Figure 3 are a symptom of the above quoted fluctuations. They serve also to give an estimation of the uncertainty of our measures, of the order of some few percents.

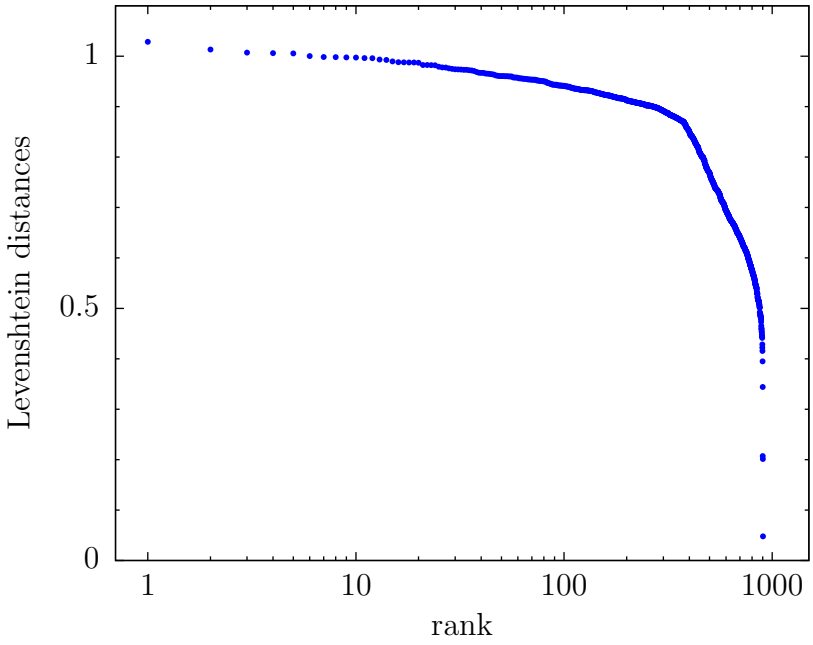

Figure 3: Ranking plot for the 903 Levenshtein distances measured for each pair of 43 languages belonging to the Tupian family. Not only this particular case, all 89 real families we use in this work present similar ranking plots, with the clear elbow observed here. These families have more than 3 languages, but their sizes varies up to the Austronesian family with its 706 languages, 248,865 pairs.

In order to compare the theoretical ranking plot like Figure 2, with a real

counterpart like Figure 3, we did a computer simulation. An evolutionary tree is dynamically constructed starting from a single founder language assigned with a given bitstring (for instance $00000\ldots0$, or any other). At each time step this bitstring suffers a random mutation, i.e. a randomly chosen bit is flipped from 0 to 1 or vice-versa. Also with a small probability at each time step, a bifurcation can occur. In this case, the current bitstring is copied to the new branch which then will evolve by itself according to its own future randomness. The number of branches (or languages) grows. At a certain moment (today) we stop the simulation and verify the bitsrings corresponding to each current, alive language. They are different from each other. We can measure the (Hamming) distance of a pair simply by counting how many bits along one bitstring differ from the corresponding bit along the other. With this simulation, contrary to reality, we have access also to the whole historical record, we know when each past bifurcation have occurred and all its descent.

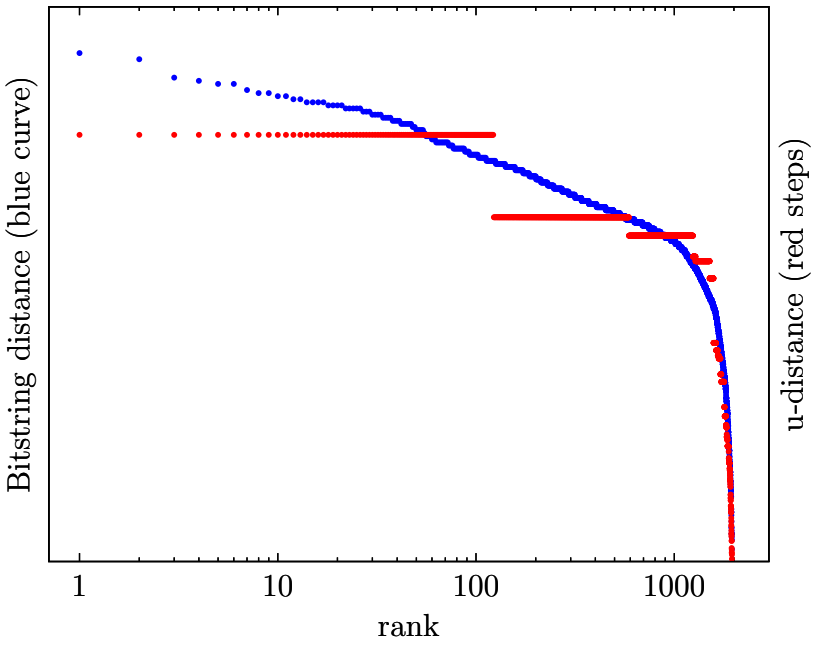

Figure 4: Computer-made model tree with 63 alive languages, 1953 pairs. The elbow is located at the same rank region of the two secondary plateaux (genera birth).

Therefore, we are able to superimpose two ranking plots, one for the known historical u-distances of each bifurcation (including their degeneracies), and the other for the bitstring distances measured only among the languages alive at the last configuration, at the moment when the simulation stopped. Figure 4 shows the result for a family with 63 languages, 1953 pairs at the end of the simulation. The uppermost plateau corresponds to the whole family, two other secondary plateaux define two genera. The vertical scales matches to each other by choosing the height of the first plateau to coincide with the average of the corresponding measured bitstrings distances, i.e. the average of the leftmost blue points in Figure 4, in the same total amount as the first plateau. Anyway, the vertical scales are irrelevant, thus omitted. The important point is that the rank corresponding to the elbow in the blue curve is located at the same region where genera appear, the two secondary red steps, independent of the vertical scales. Other simulations generating different trees give the same result.

Now, in a plot like Figure 3 for each real language family, we can measure two heights, F and G shown in Figure 2. The procedure follows. First, we determine the elbow position where the product rank $\times$ distance is maximum, getting the elbow rank E. It is the accumulated-width of the 3 first plateaux. The width of the uppermost plateau alone is $n(N-n)$, where N is the known total number of languages, and n is the unknown number of languages belonging to the smallest main branch ($N-n$ belong to the other). We assume the same ratio n/N for the next two bifurcations, allowing to determine their widths as functions of n. Finally, we equate the sum of these three widths to E, an equation whose solution provides the value of n. Therefore, the width $n(N-n)$ of the uppermost plateau becomes known, and we can average the $n(N-n)$ largest measured distances obtaining F. For G, we average the next $E-n(N-n)$ distances.

The assumption of the same ratio n/N for these bifurcations is a rule of thumb to obtain n, and needs verification a posteriori, i.e. to compare the resulting n with other possible neighboring values of n. We verified that F and G are almost insensitive to changes in n, except for $n=1$, as expected because this is just the limit of largest dependence of the width $n(N-n)$. That is why we did not include families with only 3 languages in our data. Also, for few other families for which the result of our procedure is $n=1$, we also discard the isolated branch (the Ainu family with 24 languages is an example, treated here with only 23 languages). Moreover, to consider this single isolated language within a family as a whole genus is anyway a

problematic interpretation, and thus we prefer to discard it. In these few cases, the largest branch plays the role of the family.

The final result is in Figure 5, one point per language family.

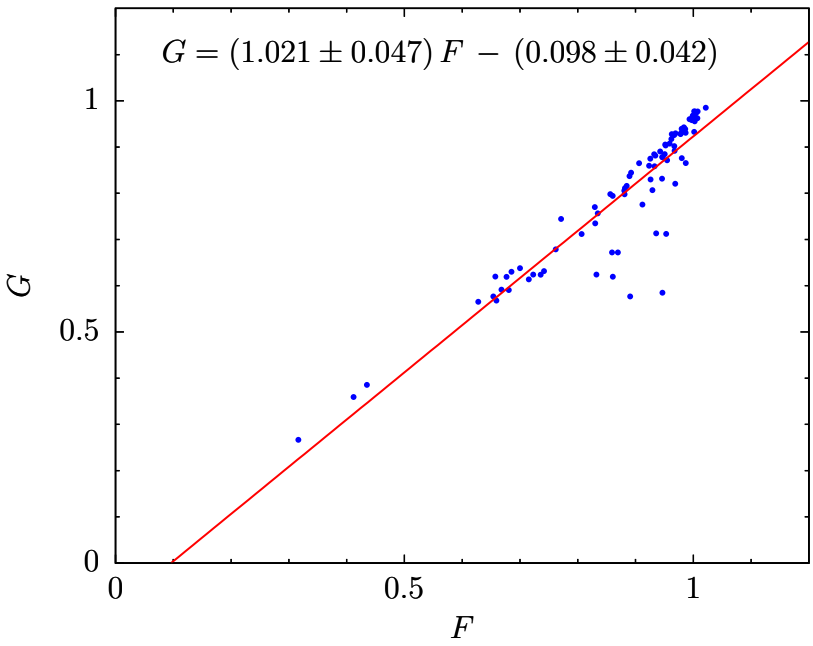

Figure 5: Genera age versus Family age for 89 families. The clear linearity, with unitary slope, indicates a universal rate of evolution for all the set. See text.

As one can see, the data are very dispersed, as denounced by the linear constant 0.098 ± 0.042, with almost fifty percent uncertainty. It is a social system, fluctuations are important. However, the same data is clearly positioned along a straight line. More than that, the angular coefficient 1.021 ± 0.047 presents a much smaller uncertainty of only few percent, in agreement with our earlier estimation. It is possible by the procedure of accumulating many families, 89 in our case.

Let's assume this angular coefficient as unitary, in agreement with its own error bar. Then, we can write the equation in Figure 5 as

$$F - G = 0.098 \pm 0.042$$

meaning that families bifurcate in genera after a constant amount of time $F - G$. The reader may wonder that this "constant" is almost fifty percent uncertain! Right! For a sampled family. That is why we don't try to translate this number in millennia[1]. For the whole set, statistically, the uncertainty is less than 5 percent, according to our error bar for the angular coefficient.

In short, our result is twofold. From the angular coefficient, Figure 5, we conclude that languages evolve according to the *same* rate, within a reasonable certainty (5% uncertainty). From the linear constant, we have only a poor estimation for *this* value, within 50% uncertainty. This result indicates a constant (although under strong fluctuations) rate of language evolution, i.e. a universal behaviour probably depending only on the human speaking capacity, not on the particular structure of the spoken language or family.

References

[1] M. Swadesh, *International Journal of American Linguistics* **21**, 121 (1955).

[2] V. Levenshtein, *Cybernetics and Control Theory* **10**, 707 (1966).

[3] E.W. Holman, S. Wichmann, C.H. Brown, V. Velupillai, A. Müller and D. Bakker, *Folia Linguistica* **42**, 331 (2008); E.W. Holman, S. Wichmann, C.H. Brown, V. Velupillai, A. Müller, D. Bakker, in: A. Arppe, K. Sinnemäki, U. Nikanne (Eds.), *Quantitative Investigations in Theoretical Linguistics*, University of Helsinki, Helsinki, 2008, pp. 4043.

[4] S. Wichmann, E.W. Holman, D. Bakker and C.H. Brown, *Physica* **A389**, 3632 (2010).

[1] Moreover, the calibration formula is not linear, and cannot be applied directly to the difference $F - G$, but to F and G separately, which makes even worse the accuracy when these values are near unity, i.e. large times.