American Journal of Systems and Software, 2016, Vol. 4, No. 2, 32-39
Available online at http://pubs.sciepub.com/ajss/4/2/1
© Science and Education Publishing
DOI:10.12691/ajss-4-2-1

Diversity for Texts Builds in Language L(M$_T$) II: Indexes Based in Abundances

José Luis Usó-Doménech[1], Josué-Antonio Nescolarde-Selva[1,2,*], Miguel Lloret-Climent[1], Meng Fan[2]

[1]Department of Applied Mathematics, University of Alicante, Alicante, Spain
[2]School of Mathematics and Statistics, Northeast Normal University, Changchun, China
*Corresponding author: josue.selva@ua.es

Abstract One saw previously that indications of diversity I_T and the one of Shannon permits to characterize globally by only one number one fundamental aspects of the text structure. However a more precise knowledge of this structure requires specific abundance distributions and the use, to represent this one, of a suitable mathematical model. Among the numerous models that would be either susceptible to be proposed, the only one that present a real convenient interest are simplest. One will limit itself to study applied three of it to the language L(M$_T$): the log-linear, the log-normal and Mac Arthur's models very used for the calculation of the diversity of the species of ecosystems, and used, we believe that for the first time, in the calculation of the diversity of a text written in a certain language, in our case L(M$_T$). One will show advantages and inconveniences of each of these model types, methods permitting to adjust them to text data and in short tests that permit to decide if this adjustment is acceptable.

Keywords: *abundances, complex systems, distribution, language, law of Zipf, model, probability, text*

Cite This Article: José Luis Usó-Doménech, Josué-Antonio Nescolarde-Selva, Miguel Lloret-Climent, and Meng Fan, "Diversity for Texts Builds in Language L(M$_T$) II: Indexes Based in Abundances." *American Journal of Systems and Software*, vol. 4, no. 2 (2016): 32-39. doi: 10.12691/ajss-4-2-1.

1. Introduction

The models that we propose for complex systems (ecological and social systems) are those based on the Dynamic of Systems [1] with the modifications made by the authors [2], with which it becomes clear that we do not expect to create a generically theory of models, but a specific form, as we consider them one of the most generalized and possibly most powerful among the wide range of alternatives offered to the modeler. For this special type of models the authors have built a language which they have called L(M$_T$) ([3-12]) whose syntax is, on a wide scale, the following:

1) The *primitive monoad* or *alphabet* A is formed by a set W of characters used to express measurable attributes $W=\left\{w_1,w_2,...,w_{n,...}\right\}$, a set D of differential functions in relation to time $D=\left\{\dfrac{d}{dt}\right\}$ and a set Φ of n-order monoads $\Phi=\left\{\left\{\varphi^1\right\},\left\{\varphi^2\right\},...,\left\{\varphi^n\right\}\right\}$. The W set is formed by the input and state variables, and $A=W\cup D\cup\Phi$.

2) The *textual alphabet* A_t is jointly built with the alphabet A and the set R of real numbers (model parameters) $R=\left\{r/r\in\Re\right\}$.

3) The *Simple Lexical Units* (SLUN) are constituted by the elements of the set A-D.

4) The *Operating Lexical Units* or operator-LUN (op-LUN) are the mathematical signs +, -.

5) The *Ordenating Lexical Units* or Ordenating-LUN (or-LUN) are the signs =, <, >.

6) The *Special Lexical Unit* (SpLUN) is the sign d/dt, which belongs to the alphabet A and defines the beginning of a phrase (state equation). The *differential vocabulary* or d-vocabulary of a measurable attribute w, V_w^{∂} , is the set formed by all partial derivatives of any order of w with respect to any other measurable attribute and the time t.

7) The *primary differential vocabulary*, $V_w^{1\partial}$, is the set formed by all partial derivatives of order 1 of w with respect to any other measurable attribute and the time t. $V_w^{1\partial}=\left\{\dfrac{\partial w}{\partial t},\dfrac{\partial w}{\partial y},...\right\}$.

8) Secondary a *higher order differential vocabularies* may also be defined and will be denoted by $V_w^{n\partial}$, $n\geq 1$. For ease of calculation in practical complex system modeling, we define a subset of $V_w^{1\partial}$ called *dimensional primary differential vocabulary*, ${}^{XYZt}V_w^{1\partial}$, consisting of all partial first order derivatives of the measurable attribute w with respect to the three spatial dimensions X, Y, Z and time t, ${}^{XYZt}V_w^{1\partial}=\left\{\dfrac{\partial w}{\partial X},\dfrac{\partial w}{\partial Y},\dfrac{\partial w}{\partial Z},\dfrac{\partial w}{\partial t}\right\}$.

To implement the models of the System Dynamics [1], a subset of cardinal 1, ${}^{t}V_w^{1\partial}$, and whose only element is the partial derivative of the p-symbol with respect to the time, will be used.

9) Let $w_1, w_2, ..., w_n$ be a set of measurable attributes. The *differential Lexicon*, d-L, is the set formed by the d-vocabularies generated by the measurable attributes,

$$d-L=\left\{\begin{matrix} V_{w_1}^{1\partial}, V_{w_2}^{2\partial}, ..., V_{w_2}^{n\partial}; V_{w_2}^{1\partial}, V_{w_2}^{2\partial}, ..., \\ V_{w_2}^{n\partial}; ...; V_{w_n}^{1\partial}, ..., V_{w_n}^{n\partial} \end{matrix}\right\}.$$

10) The Elements of d-L will be called *d-symbols*. The characters (,), { ,}, [,], are simply signs since they lack of meaning and they are the equivalent to the signs ?, !, ; (,) in the natural languages.

11) The *Separating of Lexical Units* (s-LUN) are the signs * and /.

12) The *Composed Lexical Units* (CLUN) are the strings of a SLUN separated by a s-LUN.

13) The *syllables* or composed Lexical units (CLUN) are constituted by a SLUN, or a chain of them, separated by an op-LUN or a or-LUN.

14) The *word* is the SLUN or CLUN. The symbols [·] preceding the other symbols + or – are word separations.

15) The *opsep vocabulary* V^S is the one formed by operating and separating LUNs. $\otimes \in V^S$; $\otimes = \{+, -, *, :\}$ and it will be written a element of *VS* by $\otimes$.

16) A *simple sentence* is a flow variable [1]. It is built by a CLUN or a combination of CLUNs.

2. Distributions of Abundances

Suppose known strengths of symbols in a text T built in L(M$_T$), with Q$_T$ the number of different symbols and N the number of symbols in T. The distribution of abundances is the distribution of frequencies while sequencing symbols by order of decreasing absolute or relative frequencies. The graphic representation of such a distribution of abundances will make carrying in abscises the rank $r_i; i \in (1, N)$, and in ordinates the corresponding frequencies $f(s_i)$. A particular case of this distribution is the law of Zipf [13]. The authors have treated the application of this law for the language L(M$_T$) in precedent works. In the theoretical case where touts symbols would even have frequencies (maximal diversity $\log_2 N$ and equitability 1) all not representative points would be found in a horizontal of ordinate $\frac{\sum f(s_i)}{N}$. In fact, the polygons as joining points comes closer more or less of a curve in J reversed, whose concavity is as much more accused that the equitability is weaker. It comes because there are less symbols whose strengths are superior to the average, that has some whose strengths are lower to this average there, or again that the very abundant symbols are less numerous than the rare symbols. To palliate inconveniences of this asymmetry, one often uses a logarithmic scale for ordinates and as sometimes for abscises. One can have therefore of diagrams of $r_i, f(s_i)$, $r_i, \log f(s_i)$ and $\log r_i, \log f(s_i)$.

3. Log-Linear Distribution

The simpler model is the one where logarithms of frequencies are aligned on a right of slope a:

$$\log f(s_i) = ar_i + b \tag{1}$$

and while doing $r_i = 1$

$$\log f(s_1) = a + b \tag{2}$$

The model can write itself therefore

$$\log f(s_i) = a(r_i - 1) + \log f(s_1) \tag{3}$$

As putting $a = \log \tau$ the previous relation is equivalent to:

$$f(s_i) = f(s_1)\tau^{r_i - 1} \tag{4}$$

Frequencies form a geometric progression of reason τ:

$$\begin{aligned} &f(s_1) \\ &f(s_2) = f(s_1)\tau \\ &f(s_3) = f(s_1)\tau^2 \\ &.......................... \\ &f(s_{Q_T}) = f(s_1)\tau^{Q_T - 1} \end{aligned}$$

and $N = \sum_{i=1}^{Q_T} f(s_i) = f(s_1)\left[1 + \tau + \tau^2 + ... + \tau^{Q_T - 1}\right]$.

While multiplying by τ the two members of the last equality, it comes:

$$\tau N = f(s_1)\left[1 + \tau + \tau^2 + ... + \tau^{Q_T}\right] \tag{5}$$

from where

$$N(1-\tau) = f(s_1)\left(1 - \tau^{Q_T}\right) \tag{6}$$

and

$$N = f(s_1)\frac{\left(1 - \tau^{Q_T}\right)}{1 - \tau} \tag{7}$$

The number τ can be called *constant of text*. It is the antilogarithm of the slope of the right. More the diversity of text is weak, more the slope is strong in absolute value because, frequencies having been arranged in decreasing order, the slope of the right is always negative, what comes back to say τ is always lower to the unit. A geometric progression is determined entirely by the value of the parameter τ . Indeed, the relative frequency distribution doesn't change if one either multiplies or divides all frequencies by an any number and in particular by $f(s_1)$. The model cuts down then to $\log f(s_i) = (r_i - 1)\log \tau$. To adjust the model to a distribution of abundances will come back to calculate the slope of the regression right therefore of $\log f(s_i)$ in r_i. If points are sufficiently well aligned, the right that represents the adjusted model is drawn directly and its slope gives the value of the constant of text τ. This right passes inevitably by the point having for ordinate the average of frequency logarithms and for abscissa the average of ranks equals to $\frac{Q_T+1}{2}$. If points are not aligned well either if one wants a bigger precision, one calculates the equation of the regression right of $\log f(s_i)$ in r_i. This right passes by the point of abcisse $\frac{Q_T+1}{2}$ and ordinate $\frac{\sum_{i=1}^{Q_T}\log f(s_i)}{Q_T}$. The models of this class have been used in Ecology by Utida [14] and Motomura [15].

4. Log-Normal Distribution

A log-normal model is a model in which logarithms of frequencies are distributed at random around their average $m=\frac{\sum_{i=1}^{Q_T}\log f(s_i)}{Q_T}$. This normal distribution is represented by an equation of the shape:

$$y=\frac{Q_T+1}{\sigma\sqrt{2\pi}}e^{\left[-\frac{(\log f(s)-m)^2}{2\sigma^2}\right]} \qquad (8)$$

and as taking like origin dawns it means, that is while putting $\log f(s)-m=R$

$$y=\frac{Q_T+1}{\sigma\sqrt{2\pi}}e^{\left[-\frac{R^2}{2\sigma^2}\right]} \qquad (9)$$

It is convenient to use logarithms of preference basis 2 to vulgar logarithms (Preston, [16], [17]). The corresponding interval to a R unit is then the octave, that is, the interval in which the frequency doubles of value. When one passes an octave to the superior octave, R increases a unit. The surface $\int_{R_1}^{R_2} ydR$ understood between the axis of abscissas, the curve and the two ordinates R_1,R_2 represents the number of symbols of which frequencies its understood between $f(s_1), f(s_2)$ as $\log_2 f(s_1)-m=R_1, \log_2 f(s_2)-m=R_2$. The normal curve represented by the equation (9) is the *curve of symbols*. It is defined mathematically for all values of R, of $-\infty$ to $+\infty$ and the total surface $\int_{-\infty}^{+\infty} ydR$ is equal to Q_T+1. However, extremities of the two branches of the curve, asymptotes to axis of R, cannot have linguistic significance already that the useful part some curve is limited to the number of octaves really covered by the distribution of abundances observed in the text. Of a theoretical point of view, one will admit that boundary-marks of this useful interval must be symmetrical in relation to the origin. Besides, one will admit that the position of these boundary-marks $-R_M,+R_M$, is as the area understood between the curve, the axis of the R and outside to the useful interval either equal to a symbol. It comes back to say that the useful surface is equal to Q_T, number of symbols really existing in the text and that one disregards the useful surface on all sides, an equal area to 0.5. It results some that is defined by the integral:

$$\frac{Q_T+1}{\sigma\sqrt{2\pi}}\int_{-R_M}^{+R_M} e^{\left[-\frac{R^2}{2\sigma^2}\right]}dR=Q_T \qquad (10)$$

While putting $R=t\sigma, R_M=x\sigma, dR=\sigma dt$ and the formula (10) becomes

$$\frac{1}{\sqrt{2\pi}}\int_{-x}^{+x} e^{\left[-\frac{t^2}{2}\right]}dt=\frac{Q_T}{Q_T+1}=\Theta(x) \qquad (11)$$

Therefore, the theoretical boundary-marks of useful interval for the curve of symbols are equal to the standard deviation σ multiplied by $\pm x$, x being read in tables of function Θ relative to reduced normal curve

$$\Theta(x)=\frac{Q_T}{Q_T+1} \qquad (12)$$

Let's consider the following series $R_1,R_2,...,R_{Q_T}$, understood inside the useful interval $(-R_M,+R_M)$ and definite according to frequencies of symbols by $R_1=\log_2 f(s_1)-m, R_2=\log_2 f(s_2)-m,...,R_{Q_T}$ $=\log_2 f\left(s_{Q_T}\right)-m$. The useful surface Q_T, can be divided in many partial surfaces equal each one to unity, is to say, correspondent each one to a symbol, and limited by parallels to the axis of ordinates framing values $R_1,R_2,...,R_{Q_T}$. The two extreme parallels are evidently those of boundary-marks $-R_M,+R_M$. If are cumulated these partial surfaces, one gets the series Q_T of the first whole numbers 1, 2,..., Q_T.The curve integral of Gauss that represents the increase of the surface accumulated according to R, is other that the curve of variation of r_i according to $\log_2 f(s_i)-m$. It passes therefore by points of co-ordinates $(r_i,\log_2 f(s_i)-m)$. This curve has a characteristic sigmoid shape. It is symmetrical in relation to the average of ranks and in relation to the average of frequency logarithms. A log-normal abundance distribution on a diagram in r_i and $\log_2 f(s_i)$, deal therefore an integral curve of Gauss. These curves are transformed in rights of probits, when one replaces the accumulated surfaces expressed in percentages by their probits $P(k_i)$. However, to calculate these percentages, it is necessary to take account the total surface understood between the normal curve and the axis of abscises of $-\infty$ to $+\infty$. This total surface is equal to Q_T+1. And beyond of each boundary-marks of the useful interval, the surface is equal to 0.5. It agrees to take like percentage $\frac{Q_T+0.5}{Q_T+1}$

for the superior boundary-mark of the useful interval and $\frac{0.5}{Q_T+1}$ for the lower boundary-mark.

The right of probits passes by the point having for co-ordinates P(0.5) = 5 and $\log_2 f(s) = m$. Its slope is equal to σ. Its equation is the shape:

$$\sigma\left[P(k_i)-5\right] = \log_2 f(s_i) - m \tag{13}$$

This equation permits to calculate the theoretical strengths of a distribution log-normal by the three parameters σ, Q_T, m.

In the curve of symbols each of octaves, correspond a certain number of symbols whose frequency is roughly the event. There is a median or modal octave whose middle frequency $f(s_0)$ has for logarithm the average of logarithms of frequencies $\log f(s_0) = m$. This octave has for limits m - 0.5 and m + 0.5. Let y_0 be the number of symbols whose frequency falls in this octave; y_0 is the ordinate to the top of the symbols curve. It is therefore equal to $\frac{Q_T+1}{\sigma\sqrt{2\pi}} = 0.3989\frac{Q_T+1}{\sigma}$. The symbols number corresponding to the modal octave is therefore $y_0 f(s_0)$. One gets the number of symbols corresponding to every octave in the same way while multiplying the number of the different symbols by the middle frequency of symbols. For the octave of rank R, the average frequency is $f(s_0)2^R$ and the number of different symbols is $y = y_0 e^{\left[-\frac{R^2}{2\sigma^2}\right]}$. While multiplying the two numbers, one gets a new distribution and a new curve, so-*called curve of individual symbols*, because its equation permits to calculate the number of individual symbols corresponding to one interval (R_1, R_2), that is, whose frequencies are understood between $f(s_1), f(s_2)$ as $R_1 = \log_2 f(s_1) - m, R_2 = \log_2 f(s_2) - m$. This curve is also a normal curve. Its equation is the following:

$$Y = f(s_0)2^R y_0 e^{\left[-\frac{R^2}{2\sigma^2}\right]} \tag{14}$$

Under this shape, the properties of individual symbols curve don't clearly appear. It agrees to transform its equation. One first replaces 2^R by $e^{[R\log 2]}$ what permits to write, in grouping the 2 exponential terms in one alone:

$$Y = f(s_0) y_0 e^{\left[R\log 2 - \frac{R^2}{2\sigma^2}\right]} \tag{15}$$

$$Y = f(s_0) y_0 e^{\left[\frac{2\sigma^2 R\log 2 - R^2}{2\sigma^2}\right]} \tag{16}$$

One transforms the numerator of the expression then between hooks of way to regroup all terms in R^2 and R in only one square. It is sufficient to write:

$$\begin{aligned} &2\sigma^2 R\log 2 - R^2 \\ &= \left(\sigma^2\log 2\right)^2 - \left(\sigma^2\log 2\right)^2 + 2\sigma^2\log 2 - R^2 \\ &= \left(\sigma^2\log 2\right)^2 - \left(R - \sigma^2\log 2\right)^2 \end{aligned}$$

Therefore

$$Y = f(s_0) y_0 e^{\left[\frac{\left(\sigma^2 R\log 2\right)^2}{2\sigma^2} - \frac{\left(R-\sigma^2\log 2\right)^2}{2\sigma^2}\right]} \tag{17}$$

while putting

$$f(s_0) y_0 e^{\left[\frac{\left(\sigma^2 R\log 2\right)^2}{2\sigma^2}\right]} = Y_0 \tag{18}$$

finally:

$$Y = Y_0 e^{\left[-\frac{\left(R-\sigma^2\log 2\right)^2}{2\sigma^2}\right]} \tag{19}$$

One recognises the equation of a normal curve of which the ordinate to the top is Y_0, whose standard deviation is σ and whose average is $\sigma^2\log 2$. The top of the individual symbols curve is baffled of $\sigma^2\log 2$ in relation to the curve of symbols. The top of the individual symbols curve occupies therefore, in relation to the extremity of the useful part of symbols curve, a position that depends of σ. Preston [17] called *canonical distributions* those for the extremity of the useful part of curve of symbols and the top of the individual symbols curve have the same abscise. One then the equality $\sigma^2\log 2 = x\sigma$ from where $\sigma = 1.443x$. It results that to a value data of Q_T it corresponds a canonical distribution of which all parameters are determined. It is called *constant of the text of Preston* m' the inverse of the square of the standard deviation $m' = \frac{1}{\sigma^2}$. When the graphic determination of m and is difficult or σ is not sufficiently precise, it is always possible to resort to the equation of the probits right $\sigma(P-5) = \log f(s) - m$ or $\log f(s) = \sigma P + b$ while putting $b = m - 5\sigma$. One takes like equation the one of the regression right of $\log f(s_i)$ in $P(k_i)$.

The method of Preston has been used in Ecology.

5. Distribution of Mac Arthur

In the distribution of Mac Arthur ([18,19,20]), the frequency of symbol of rank r_i from most abundant is given by the formula:

$$f(s_i) = \frac{N}{Q_T}\sum_{l=1}^{l=Q_T+1-r_i}\frac{1}{Q_T-l+1} \tag{20}$$

The more abundant symbol has for frequency:

$$f\left(s_1\right)=\frac{N}{Q_T}\left[\frac{1}{Q_T}+\frac{1}{Q_T-1}+\ldots+\frac{1}{1}\right] \quad (21)$$

and the rarer symbol

$$f\left(s_{Q_T}\right)=\frac{N}{Q_T}\left[\frac{1}{Q_T}\right] \quad (22)$$

The sum of the relative frequencies is equal á 1.

$$\sum f\left(s_1\right)=\frac{1}{Q_T}\left[Q_T\frac{1}{Q_T}+\left(Q_T-1\right)\frac{1}{Q_T-1}+\ldots+\frac{1}{1}\right]$$
$$=\frac{Q_T}{Q_T}=1$$

Diagrams in r_i and $\log f\left(s_i\right)$ show that distributions of Mac Arthur are very little represented by concave and not symmetrical curves in S.

Among the studied distributions, those of Mac Arthur only depend on two parameters, N and the number of symbols Q_T, whereas those log-lineal and log-normal depend in addition to a third parameter, the constant of the text. N brought back to the volume of text is other that the density of symbols. Q_T is the specific wealth of the text. As for has the constant of text of the distribution log-lineal, whose logarithm is the slope of the right and to the constant of text of the log-normal distribution that is bound to the standard deviation of the curve of symbols and the curve of individual symbols by the relation $m'=\frac{1}{\sigma^2}$, they are closely one and the other dependent of the diversity of the text. Distributions log-lineal and log-normal are susceptible of much better adjustments that of Mac Arthur one. The comparison can take place directly between the observed and calculated frequencies. Hairston [21] and King [22] use the relation between the variance of theoretical value and the variance of observed values. The concordance will be perfect if this relation is equal to 1, and it will be less good if is more different of 1. The variance of values observed in the text is:

$$\sigma_o^2=\frac{1}{Q_T-1}\left[\sum f\left(s_i\right)-\frac{N^2}{Q_T}\right] \quad (23)$$

The theoretical value variance is:

$$\sigma_t^2=\frac{1}{Q_T-1}\left[\sum f\left(s_i\right)-\frac{N^2}{Q_T}\right] \quad (24)$$

This method is independent of the size of the text. The value of the relation $\frac{\sigma_o^2}{\sigma_t^2}$ must be considered like a simple indication encoded on the degree of concordance or conflict between the distribution observed in the text and the corresponding model of Mac Arthur to the same number of symbols Q_T. It is not necessary to assign it the same statistical significance of two sample variances that to the F of Snedecor, that calculates himself of the same way for the comparison of two sample variances, because the method of Snedecor supposes that values to leave of which are calculated variances are normally distributed: this is not the case for distributions of Mac Arthur. It is very appreciable to the size of texts and supposes that the order of symbols in the text is identical that the real order of abundances in the language, otherwise the calculated value for χ^2 would be underestimated it. Finally, it doesn't take account of the sign of gaps nor the positive and negative gap distribution. Or the positive gap (frequencies observed superior to those foreseen by the model) concerning the most often the abundant symbols and the negative gaps (frequencies observed lower to those foreseen by the model) the rare symbols. It is important to know if distribution observed for signs of gaps can or no be assigned at random. David's test ([23]) answers to this question. When the size of texts is relatively big, it often arrives that values found for χ^2, considering the number of liberty degrees, correspondent to probabilities many too weak so that one can assign at random of the sampling gaps between the observed frequencies and those foreseen by the dsitribution of Mac Arthur. In this case, even though all conditions of application of the test of χ^2 the are not rigorously full, one will be founded to conclude that the distribution of abundances in the language is not compliant to a model of Mac Arthur.

The distribution of Mac Arthur can be used especially to represent distributions of text abundances understanding few individuals and little symbols. The only problem to put would be the one of the comparison between the diversity of text and the one of the distribution. This comparison makes through the intermediary of Shannon's diversity index. The relation between the calculated index on the text and the calculated index on the model of Mac Arthur, for the same number Q_T, is equitability. The difference with the definite equitability resides in the choice of the stationary element of comparison $\log_2 Q_T$, the maximal theoretical diversity being replaced by the diversity of the model of Mac Arthur considered as the limit toward which offers the diversity of the language, being $\log_2 Q_T$ an inaccessible theoretical maximun. As index of Shannon's diversity make intervene of logarithms in their calculation, their comparison can make correctly by a simple quotient.

We proposed a new definition of the equitability, relation γ of the number Q_T of symbols observed to the number Q_T^{MA} of symbols that, distributed according to a model of Mac Arthur, would give the same indication of diversity

$$\gamma=\frac{Q_T^{MA}}{Q_T} \quad (25)$$

6. Law of Number-Frequency

Let L be the length of a text, where L is the number of signs of the same.

Let p(r) be the probability of occurrence of a sign w of rank r, then the probability that the sign w of rank r appear i sometimes in the text is:

$$p\left[n_i(r,L)\right]=\binom{i}{L}p(r)^i\left[1-p(r)\right]^{L-i} \tag{26}$$

We define the random variable $N_i(L)$, number of occurrences of the sign of rank i in the text of length L as:

$$N_i(l)=\sum_r n_i(r,L) \tag{27}$$

being $n_i(r, L) = 1$ if the sign w appears, and $n_i(r, L) = 0$, if it does not. These variables $n_i(r, L)$ they are not independent, but the average of the sum is the sum of the averages, then:

$$\begin{aligned}E\left[N_i(L)\right]&=\sum_r E\left[n_i(r,L)\right]\\&=\sum_r n_i(r,L)\,p\left[n_i(r,L)\right]=\binom{i}{L}\sum_r p(r)^i\left[1-p(r)\right]^{L-i}\end{aligned} \tag{28}$$

Considering that

$$p(r)=P\cdot r^{-\beta}=P\cdot r^{-\frac{1}{\Theta}} \tag{29}$$

and by p(r) = x, it follows that

$$r=x^{-\Theta}p^{\Theta} \tag{30}$$

For large values of L, the above sum differs little from the sum, restricted to some range, such as (10, ∞) and the following integral:

$$\begin{aligned}&\int_0^1 x^i(1-x)^{L-i}\,dr(x)=P^T T\int_0^1 x^{i-T-1}(1-x)^{L-i}\,dx\\&=P^T T\frac{\Gamma(i-T)\Gamma(L-i-1)}{\Gamma(T-L-1)}\end{aligned} \tag{31}$$

differs little from the integral restricted to (0, p (10)), finally the restricted sum, and the restricted integral differ little, thus:

$$\begin{aligned}E\left[N_i(L)\right]&\cong\frac{L!}{(L-i)!i!}P^T T\frac{\Gamma(i-T)\Gamma(L-i-1)}{\Gamma(T-L-1)}\\&=\frac{\Gamma(L+1)}{\Gamma(i+1)}P^T T\frac{\Gamma(i-T)}{\Gamma(L+1-T)}\\&=P^T T\frac{\Gamma(i-T)}{\Gamma(i+1)}\frac{\Gamma(L+1)}{\Gamma(L+1-T)}\end{aligned} \tag{32}$$

For large enough L it is:

$$E\left[N_i(L)\right]\cong P^T TL^T\frac{\Gamma(i-T)}{\Gamma(i+1)} \tag{33}$$

and to i big:

$$E\left[N_i(L)\right]\cong P^T TL^T i^{-(T+1)} \tag{34}$$

which is the expression of law number-frequency.

7. Lexic Unities Model

Lexical units can be grouped according to different criteria:

1. According to the type of behavior they describe. The same behavior can be described by different lexical items, which we will call synonymous; we can then form groups of synonyms LUN.
2. According to the functions contained therein. For example we can group the periodic functions, or those whose power series developments are similar.
3. According to their probabilities of occurrence.
4. According to the number of primitive symbols that compose them.

We are interested in estimating the number of lexical units that belong to a certain group and containing a certain amount of primitive symbols.

Let N be lexical units are classified into k non-empty classes. Let N_i be the number of lexical units in the ith class. These N_i lexical units in the ith class are divided in Mi symbols according to a Bose-Einstein distribution, ie:

$$\mathrm{P}\left[\mathrm{L}=\left(\mathrm{l}_1,\ \mathrm{l}_2,\ \ldots\ldots\mathrm{l}_{\mathrm{M}}\right)\right]=\binom{N-1}{M-1}^{-1} \tag{35}$$

with $l_i \geq 1$ and $\Sigma\, l_i = N$.

Let $G_i(S)$ be the number of lexical units that are, with exactly s symbols in the i-th class, then it is:

$$G(s)=\sum_{i=1}^{k}G_i(S) \tag{36}$$

$$M=\sum_{i=1}^{k}M_i \tag{37}$$

The proportion of lexical units with exactly s symbols is:

$$\frac{G(S)}{M}=\frac{\sum_{i=1}^{k}G_i(S)}{M}=\sum_{i=1}^{k}\left(\frac{M_i}{M}\right)\frac{G_i(S)}{M_i} \tag{38}$$

Considering only one class, for example the first, we find that:

$$\frac{G_1}{M_1}=\theta_1(1-\theta_1)^{S-1}+\delta_{n_1} \tag{39}$$

when $N_1 \to \infty$, where $\theta_1 = M_1/ N_1$, and δ_{N1} it converges in probability to 0. Similarly, it states:

$$\frac{G_i(S)}{M_i}=\theta_i(1-\theta_i)^{S-1}+\delta_{n_i} \tag{40}$$

for each i, and furthermore, the weighted average G(S)/M (38) converges to a constant p(s):

$$p(S)=\int_0^1 t(1-t)^{S-1}\,dH(t) \tag{41}$$

for $S \geq 1$ wherein H is a function of proper distribution. This constant for $S \to \infty$ and a variety of possible H becomes:

$$p(S)\approx CS^{-(1+\alpha)} \tag{42}$$

where:

$$\frac{G(S)}{M}\approx C.S^{-(1+\alpha)} \tag{43}$$

Whereupon under this model, the individual occurrences of G(S)/M can be expected to follow approximately the law of Zipf.

8. An Application of the Law of Zipf

For simplicity let us first consider that there are only four signs in the text, that the energy of any of the signs is restricted to one of the values E(x) = 0, ΔE(x), 2ΔE(x), 3ΔE(x) and that the total energy of the text is 3ΔE(x).

Since the signs can exchange energy between them, all divisions of the total energy between 4 signs are possible. Consider the possible cases:

I) Three signs have E (x) = 0 and one has E(x) = 3ΔE(x). There are four different ways to achieve this division of energy as any of the 4 signs can be found in the energy state 3ΔE(x). That is, the number of duplicate distinguishable divisions is 4r.

II) Two signs have E(x) = 0, the third E(x) = ΔE(x) and fourth E(x) = 3ΔE(x). In this case there are 12 different ways to achieve this division. Ie the number of distinct divisions duplicate is 12.

III) One sign is E(x) = 0, and the remaining three have E(x) = 3ΔE(x), there are 4 different ways to achieve this scheme. Ie the number of duplicate distinct divisions is 4.

In assessing the number of duplicate divisions is counted as distinct duplicate, any arrangement of signs between different energy states. However, any new arrangement of signs in the same state of energy are not counted as duplicates, because equals signs, and have the same energy, cannot be distinguish one from another. That is, identical signs are treated as if they were distinguishable, except for new arrangements in the same energy state.

The total number of permutations of the 4 signs is 4!. If we consider n signs, the number of different orderings is n!, but new arrangements within the same energy level do not count, for example in the case II) the number of distinct divisions is reduced from 4! To 4!/2!. That is 12, since there 2! Arrangements in the state E (x) = 0 do not count as distinguishable. For cases I) and III), the divisions are reduced from 4! to 4!/3! That is 4, and that there are 3!.

New arrangements in the state E(x) = 0, or the state E(x) = ΔE(x) do not count as distinguishable. Since all possible divisions of energy occur with the same probability, the probability that a given type divisions occur it is proportional to the number of duplicate distinguishable divisions from this type, and then the probability P_i is exactly equal to this number divided by the total number of those divisions. So the probabilities for the three cases considered are: 4/20, 12/20 and 4/20.

Let's see what the probable number, N´(E(X)) of signs in the state energy E(x).

a) In the state of energy E(x) = 0for the case I) there are 3 signs in this state and the probability P_i = 4/20.

b) For the case II) are two signs in this state and P_i = 12/20.

c) For the case III) there is a sign and Pi = 4/20, then the likely number of signs with zero energy, N'(0) is N´(0) = 3 (4/20) + 2 (12/20) + 1 (4/20) = 2

For the remaining cases are: N´(ΔE(x)) = 24/20, N´(2ΔE(x)) = 12/20 and N´(3ΔE(x)) =4/20 and N´(4ΔE(x)) = 0 , which they add 4 as expected.

9. Conclusions

From all the above we can draw the following conclusions:

1. The relationship empirically observed for the media, or Zipf distribution of values is an example of a distribution that satisfies this constraint: the power-law distributions are stable under charted.
2. The structure of Zipf exactly fulfills the restriction to increase the structure, and the identifiability of features.
3. Zipf structure can evolve by observing basic mechanisms that favor the structure involved in complex systems.

The appearance of Zipf's law is not caused by accident; it can be understood at a level which considers the interaction between the laws of complex systems, which, unlike the laws of physics can be developed. In general one can speak of coevolution of laws (behavior) and objects. We believe that to better understand the complex systems is necessary to take into account these interrelationships.

References

[1] Forrester. J.W. 1961. *Industrial Dynamics*. MIT Press, Cambridge, MA.

[2] Usó-Domènech, J. L., Mateu, J and J.A. Lopez. 1997. Mathematical and Statistical formulation of an ecological model with applications. *Ecological Modelling*. 101, 27-40.

[3] Nescolarde-Selva, J.; Usó-Doménech, J. L.; Lloret-Climent, M. 2014. Introduction to coding theory for flow equations of complex systems models. *American Journal of Systems and Software*. 2(6). pp. 146-150.

[4] Nescolarde-Selva, J., Usó-Doménech, J.L., Lloret- Climent, M. and González-Franco, L. 2015. Chebanov law and Vakar formula in mathematical models of complex systems. *Ecological Complexity*. 21. pp. 27-33.

[5] Sastre-Vazquez, P., Usó-Domènech, J.L. and Mateu, J. 2000. Adaptation of linguistics laws to ecological models. *Kybernetes*. 29 (9/10). 1306-1323.

[6] Usó-Domènech, J.L., Sastre-Vazquez, P. and Mateu, J. 2001. Syntax and First Entropic Approximation of L(MT): A Language for Ecological Modelling. *Kybernetes*. 30(9/10). 1304-1318.

[7] Usó-Domènech, J.L. and Sastre-Vazquez, P. 2002. Semantics of L(MT): A Language for Ecological Modelling. *Kybernetes*. 31 (3/4), 561-576.

[8] Usó-Domènech, J.L., Vives Maciá, F. and Mateu. J.. 2006a. Regular grammars of L(MT): a language for ecological systems modelling (I) –part I. *Kybernetes*. 35 nº6, 837-850.

[9] Usó-Domènech, J.L., Vives Maciá, F. and Mateu. J. 2006b. Regular grammars of L(MT): a language for ecological systems modelling (II) –part II. *Kybernetes*. 35 (9/10), 1137-1150.

[10] Usó-Doménech, J. L., Nescolarde-Selva, J., Lloret-Climent, M. 2014. Saint Mathew Law and Bonini Paradox in Textual Theory of Complex Models. *American Journal of Systems and Software*.2 (4), pp. 89-93.

[11] Usó-Doménech, J. L., Nescolarde-Selva, J. 2014. Dissipation Functions of Flow Equations in Models of Complex Systems. *American Journal of Systems and Software*. 2 (4), pp. 101-107.

[12] Usó-Doménech, J. L., Nescolarde-Selva, J., Lloret-Climent, M. and González-Franco, L. 2014. Diversity for Texts Builds in Language L(MT): Indexes Based in Theory of Information. *American Journal of Systems and Software*. 2(5). pp. 113-120.

[13] Zipf, G.K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.

[14] Utida. T. 1943. Relation entre les populations expérimentales de Callosobrunchus chinensis Linné (Coléoptères) et son parasite (Hyménopttères). III. Influence de la densité de population de l'hôte sur la proliferation du parasite. *Seitaigaku Kenkyuu*, 9, 40-54.

[15] Motomura, L. 1947. Further notes on the law of geometrical progression of the population density in animal association. *Seiri Seitai*. 1, 55-60.

[16] Preston, F.W. (1948. The commoness and rarity of species. *Ecology*, 29, 254-283.

[17] Preston, F.W. 1962). The canonical dsitribution of commonness and rarity. *Ecology*, 43, 185-215 and 410-432.

[18] Mac Arthur, R.H. 1957. On the relative abundance of bird species. *Proc. Nat. Acad. Sci.* 43, 293-295.

[19] Mac Arthur, R.H. 1960. On the relative abundance of species. *Amer. Nat*., 94, 25-36.

[20] Mac Arthur, R.H. (1969. Patterns of communities in the tropics. *Biol. J. Linn. Soc.*, 1, 19-30.

[21] Hairston, N.G. 1959. Species abundance and community organization. *Ecology*. 40, 404-416.

[22] King, S.E. (1964). Relative abundance of species and Mac Arthur model. *Ecology*. 45, 716-727.

[23] David, F. N. 1947. A χ^2 smooth test for goodness of fit. *Biometrika*. 34, 299-304.