American Journal of Systems and Software, 2014, Vol. 2, No. 5, 113-120
Available online at http://pubs.sciepub.com/ajss/2/5/1

DOI:10.12691/ajss-2-5-1

Diversity for Texts Builds in Language L(M_T): Indexes Based in Theory of Information

José Luis Usó-Doménech[1], Josué-Antonio Nescolarde-Selva[1,*], Miguel Lloret-Climent[1], Lucía González-Franco[2]

[1]Department of Applied Mathematics, University of Alicante, Alicante, Spain
[2]Biodiversity Research Institute CIBIO, University of Alicante, Alicante, Spain
*Corresponding author: josue.selva@ua.es

Received September 13, 2014; Revised September 20, 2014; Accepted September 24, 2014

Abstract If one has a distribution of words (SLUNs or CLUNS) in a text written in language L(M_T), and is adjusted one of the mathematical expressions of distribution that exists in the mathematical literature, some parameter of the elected expression it can be considered as a measure of the diversity. But because the adjustment is not always perfect as usual measure; it is preferable to select an index that doesn't postulate a regularity of distribution expressible for a simple formula. The problem can be approachable statistically, without having special interest for the organization of the text. It can serve as index any monotonous function that has a minimum value when all their elements belong to the same class, that is to say, all the individuals belong to oneself symbol, and a maximum value when each element belongs to a different class, that is to say, each individual is of a different symbol. It should also gather certain conditions like they are: to be not very sensitive to the extension of the text and being invariant to certain number of operations of selection in the text. These operations can be theoretically random. The expressions that offer more advantages are those coming from the theory of the information of Shannon-Weaver. Based on them, the authors develop a theoretical study for indexes of diversity to be applied in texts built in modeling language L(M_T), although anything impedes that they can be applied to texts written in natural languages.

Keywords: *diversity, entropy, information, indetermination, language, model*

Cite This Article: José Luis Usó-Doménech, Josué-Antonio Nescolarde-Selva, Miguel Lloret-Climent, and Lucía González-Franco, "Diversity for Texts Builds in Language L(M_T): Indexes Based in Theory of Information." *American Journal of Systems and Software*, vol. 2, no. 5 (2014): 113-120. doi: 10.12691/ajss-2-5-1.

1. Introduction: The Textual Grammar

The Ecological Model as a whole, even though its methodology is not separated from the rest of the models, possesses certain characteristics that specify in its own original form the functions of these. On the one hand, they must be separated from models based on physical, mechanical causes. The existence of feedback expresses the conditions of adaptation, regulation, a structural response to equally structural signals that have been constructed on documentary bases typical of theories of biology, ecology and socio-economics. On the other hand, a special effort has to be made to overcome the ease and ambiguity of intuition. All models have in common that they encode experience and always involve signs, signals, syntaxes, semantics and an ability to decode and derive meaning from what is encoded, Gash (2014).

That is to say, it meets the conditions of a language. The authors consider the idea of language according to Chomsky (1965, 1969) as:

1. The elements are discreet and arbitrary. The only elements which are going to be relevant for the grammatical description are discreet ones.
2. Combinations of elements are linear, denumerable**.**
3. Not all combinations of sentences constitute a sentence**.** We describe a sentence (word-string) as an occurring mathematical sequence that obeys the regularities for sentence hood required by "model grammar". Not all finite sequences of elements occur as sentences. The fact that not all combinations occur makes it possible to define larger elements as restrictions on the combinations of smaller elements.
4. In language there is redundancy in respect to the sequence of ultimate elements and this redundancy is composed of a system of intermediate elements.
5. This language has a semantic meaning, the meaning of entities and the meaning of grammatical relations among them.

Obviously, if what we pretend is to build a mathematical model of a system, we will use the formal language of mathematics. Using mathematical expressions, a level of objectivity can be reached, or at least get as close to it as possible. The models that we propose are those based on the Dynamic of Systems (Forrester 1961) with the modifications made by the authors (Usó-Domènech et al.1997), 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) whose syntax is, on a wide scale, the following (Nescolarde-Selva and Usó-Doménech, 2013; Sastre-Vazquez et al., 2000; Usó-Domènech et al., 1997, 2000[a, b], 2001, 2002, 2006[a,b], 2014; Villacampa & Usó-Domènech, 1999; Villacampa et al., 1999[a, b]):

We define as *associative field of a measurable attribute* w and we called Φ_w, the set constituted by all possible symbols of said measurable attribute: $\Phi_w = \left\{\varphi_w^0, \left\{\varphi_w^1\right\}, \left\{\varphi_w^2\right\} .... \left\{\varphi_w^n\right\}, ...\right\}$. The set Φ_w will be a denumerable set. In the practical tool, it will be a requisite to define one subset $V_w \subset \Phi_w$ whose cardinal will be an integer number. The associative field of a measurable attribute w will be called *First Order Vocabulary* (FOV) or Vocabulary of order one and will be denoted by V_w^1. The elements of V_w^1 will be called t-symbols and will be denoted by $\varphi^j{}_i$, where i represents an index of the symbol and j denotes the order of transformation. The measurable attributes are a particular case of the t-symbols. The set X formed by a FOV generated by the set of measurable attributes $W = \{w_1, w_2, ...w_n\}$ will be called *Primary Lexicon* (PL) or *alphabet of the n-order monoads*, $X = \left\{V_{w_1}^1, V_{w_2}^1, ..., V_{w_n}^1\right\}$.

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 $D = \left\{\frac{d}{dt}\right\}$ 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 $A = W \cup D \cup \Phi$.

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

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

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

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

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.

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. $V_w^{1\partial} = \left\{\frac{\partial w}{\partial t}, \frac{\partial w}{\partial y}, ...\right\}$.

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, ${}^{XYZt}V_w^{1\partial} = \left\{\frac{\partial w}{\partial X}, \frac{\partial w}{\partial Y}, \frac{\partial w}{\partial Z}, \frac{\partial w}{\partial t}\right\}$.

To implement the models of the System Dynamics (Forrester, 1961), 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.

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,

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

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.

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

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

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

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

A *simple sentence* is a flow variable (Forrester, 1961). It is built by a CLUN or a combination of CLUNs.

The vocabulary of order n $V_{w_1 w_2 ... w_n}^n$ is the one formed by simple sentences

$$V_{w_1 w_2 ... w_n}^n = \begin{Bmatrix} \varphi_i \otimes \varphi_j \otimes ... \otimes \varphi_\omega; \\ \varphi_i \in V_{w_1}^1, \varphi_j \in V_{w_2}^1, ..., \varphi_\omega \in V_{w_n}^1 \end{Bmatrix}$$

$$= \left\{\Psi_{w_1 ... w_n}^n / \Psi_{w_1 ... w_n}^n = \varphi_i \otimes \varphi_j \otimes ... \otimes \varphi_\omega; \varphi_i\right\}$$

A short notation would be $\phi_{w_1, w_2, .., w_n}^n = \varphi_{i_1} \otimes ..... \otimes \varphi_{i_n}$.

The set of all vocabularies of any order is called *t-Lexicon* t-L, and it is formed by the FOV and simple sentence vocabularies.

$$t-L = \begin{Bmatrix} V_{w_1}^1, V_{w_2}^1, ..., V_{w_n}^1, V_{w_1 w_2}^2, V_{w_2 w_3}^2, ..., \\ V_{w_1 w_n}^2, V_{w_2 w_3}^2, ..., V_{w_{n-1} w_n}^2, V_{w_1 w_2 ... w_n}^n \end{Bmatrix}$$

The set Φ will be a subset of t-L.

Let $\{\phi_n\}_{i=1.,..,n} \in V_{i=1,...,n}^1$. We say that $\phi_1, \phi_2, ..., \phi_n$ are related linguistically in a *n-order relationship* and we call it $(\phi_1, \phi_2, ..., \phi_n) \in r_n$ if and only if

$(\exists \otimes \in V^S) \vee (\exists V^n_{12...n}) \vee (\exists \Psi^n_{12...n} \in V^n_{12...n})$ and $\Psi^n_{12...n} = \phi_1 \otimes .... \otimes \phi_n$. We will call R_L the whole of all linguistic relationships $r_L; L = 1, 2, ..., n$. Let $V^n_{12...n}, V^m_{12...m}, ..., V^l_{12...l}$ be vocabularies of *n, m,...,l* orders, respectively. We say that $V^n_{12...n}, V^m_{12...m}, ..., V^l_{12...l}$ are *related linguistically* and we will call it $(V^n_{12...n}, V^m_{12...m}, ..., V^l_{12...l}) \in r_V$ if and only if $V^h_{12...h} / h = n + m + ... + l$ vocabulary exists so that

$$(\exists \Psi^n_i \in V^n_{12...n}) \wedge (\exists \Psi^m_j \in V^m_{12...m}) \wedge ...$$
$$\wedge (\exists \Psi^l_k \in V^l_{12...l}) \wedge (\exists \oplus \in V^S) \wedge (\exists A^h_{ij...k} \in V^h_{12...h})$$

where $A^h_{ij...k} = \Psi^n_i \oplus \Psi^m_j \oplus ... \oplus \Psi^l_k$.

A *complex sentence* is each ordinary differential equation (ODE) or state equation, which is built by linear combination of simple sentences $A^h_{ij...k} = \Psi^n_i \oplus \Psi^m_j \oplus ... \oplus \Psi^l_k$. A *text* T = (L, A) is the concatenation of complex sentences, determined by the argument A of the text or semantic links between these complex sentences.

The *Lexicon L of a text* is the union between the t-Lexicon and the differential Lexicon, $L = t - L \cup d - L$. We can say that the text is written in a formal language, and we call it as $L(M_T)$.

Mathematical modeling of complex structural systems is the process of producing texts of mathematical relations with the rules defined by the syntax of the $L(M_T)$ with a homomorphism in respect to a conceptual semiotic system and/or ontological reality.

The past few years have seen a rapid development in novel high-throughput technologies that have created large-scale data. This data is commonly represented as networks, with nodes. A fundamental challenge to bioinformatics is how to interpret this wealth of data to elucidate the interaction of patterns and the biological characteristics. One significant purpose of this interpretation is to predict unknown functions. Although many approaches have been proposed in recent years, the challenge still remains how to reasonably and precisely measure the functional similarities to improve the prediction effectiveness (Yan, et. al., 2014; Marashi and Tefagh, 2014; Rubin, et. al., 2014; Zhu, et. al. 2010)

2. Diversity and Information in a Text-Model

Consider the lexicon L. Consider a sign system S, representing a set of texts $\{T\}$ on the lexicon L and $S = \{T\}$. By definition, the sign system S consists of all texts generated by the argument A. being A $\rightarrow$ *hypothesis* + *objective*. It is defined a *textual space* T = < A, S>. For signs of lexicon L, $\delta \in L$ there is defined a number function $E(\delta)$, which is interpreted as the *complexity of the generation of the sign* δ in the argument A. With each text $T \in S$ there is associated a *complexity of generation* E(T), equal to the sum of the complexities appearing in the sign text

$$E(T) = \sum_{\delta \in T} E(\delta) = \sum_{\delta \in V} f_\delta E(\delta) \quad (1)$$

being f_δ the number of distinct appearances of the sign δ in the text t or frequency of δ. Obviously, $\sum f_\delta = \Lambda$ or length of the text (number of equal or different signs).

2.1. Thermodynamic of Text

Using a thermodynamic analogy, E(T) will be the energetic cost or *energy of generation* of the text T. Mandelbrot (1954, 1961), propose for the Zipf's Law (1949) the following:

$$f(r) = P\Lambda(r + \rho)^{-\beta} \quad (2)$$

being ρ, β two parameters depending of the text T and $\beta > 0$, and P is determined by

$$P^{-1} = (1 + \rho)^{-\beta} + (2 + \rho)^{-\beta} + ... + (\nu + \rho)^{-\beta} \quad (3)$$

ν the number of different signs of the text T. the formula (2) can be written in probabilistic form as

$$p_r = P\Lambda(r + \rho)^{-\beta} \quad (4)$$

The parameter β is the inverse of *temperature of information* of the text T, $\Theta = \frac{1}{\beta}$. The entropy H of the text, will be determinate by Shannon's formula $H = -\sum p_r \log p_r$. $\frac{\partial H}{\partial \beta} < 0, \frac{\partial H}{\partial \Theta} > 0$. H continuously growths from 0 to $\log \Lambda$ when Θ goes from 0 to $+\infty$. H determines Θ for a given Λ. The Mandelbrot's criteria consists on transforming in 0 the variation free of $A = E(\delta) - \Theta H$, that is to say, the energy excess if the energy for symbol in the formula of Shannon was $\frac{1}{\Theta}$. A it will be the *usable energy of Helmholtz*, that is to say, the available energy for the dissipation, being $E(\delta)$ the enthalpy or *heating content of information*. Therefore it can assimilate Λ as a *text volume*, and Λ, Θ as state variables. The existence of a hypothetical text volume will make suppose the existence of a "recipient" where the components of this text exercise a hypothetical *pressure of information* P. The entropy H measures how much information lacks to understand that structure has a system that is disordered for the observer of this system. From (2)

$$\Lambda = f_r P^{-1} (r + \rho)^{\frac{1}{\Theta}} \quad (5)$$

1. The entropy H is a growing magnitude that goes of 0 to $+\infty$. Therefore in this case the information I will be 0.
2. If $\Theta = 0 \Rightarrow \Lambda = \infty$ and $H = 0$, that which is logical since the signs of very high range add very little to H or to $E(\delta)$. The information I will be 0. An infinite text is equal to an infinite volume, formed by infinite signs with a structure infinitely rigid without any

movement (appearance) of the signs. Then we will be before the *absolute zero of information*. The absolute zero of information will correspond to the maximum of information.

3. If $0 < \beta \leq 1 \Rightarrow 1 \leq \Theta < \infty \Rightarrow \Lambda \to \infty$ therefore $H \to 0$ and the information will spread to be zero. The system will spread to be more and more structured.
4. If $\Theta = \infty \Rightarrow \Lambda = f_r P^{-1} = 0$, then therefore the information will be I = 0. The structure is zero. An empty volume corresponds to the *empty text* $T = \varnothing$.
5. To take information means to make the most complex, stronger structure and logically to go bringing near the temperature of information from the system to the absolute zero of information. Contrarily, to give information means to make the weakest structure and to bring near the temperature of information to the infinite.
6. A system is *informatively colder* regarding other, when its temperature of information is more near the absolute zero of information that the other system that will be considered informatively hotter.
7. If a system informatively cold contacts another informatively hotter, the first one will cool down more, at the same time that the second system warms, increasing its temperature of information.
8. A system takes information of other when it makes more complex its structure and therefore it diminishes its temperature of information.

2.2. The Hypothesis of "soup-state"

Before the modeler prepares to create the text, a system exists constituted by the language that must use in the generation of this text, in our case the L(M_T) language. This system is constituted by all primitive symbols (n-order monoads, SLUNs) and its corresponding grammars. This way, in this initial state we can consider the existence of a singularity formed by an infinite number of elements that is in a minimum energy state and having such form that belongs together to the structure of language, but without any relationship to each other to constitute a text. In this state there will be a temperature of infinite information, a volume zero and infinite entropy. We have spoken of this singularity in the sense defined by the current cosmology (Davies, 1983), since it represents the absolute uncognisciability and where is possible to apply the Hawking's principle of absolute ignorance and it is therefore lacking of all information, that is to say I = 0 or $H = +\infty$. This singularity is in a state of maximum disorder or thermodynamic balance. If it is compared with a perfect gas, the most probable state in certain quantity of gas locked in a recipient, is the uniform density, the position of the individual molecules are at random, that is to say, any configuration among the high number of the possible ones, it will be equally probable. This would be the situation in that is our system before beginning to use the language, when it is only had this singularity to which have defined as "soup state", and is constituted by all the symbols of the language and generated by their corresponding grammar. This is the limit of our measure of information, the situation of maximum statistical entropy.

2.3. The No Selective Contraption

One of features essential of all text is its degree of organization that results notably in a certain specific abundance distribution, by a certain relative frequency specter, of the most abundant symbol to the rarest symbol. However, the relative frequency of a symbol is not other that its probability of apparition in a determined text, when the apparition is done at random by a technique or a no selective contraption. This probability is unknown. The theory of said no selective contraption is the following way:

Suppose the Nature or Ontologic System like a discreet source Δ by heart null generating data $\Delta = \{\Delta_1, \Delta_2, ..., \Delta_N\}$. This source emits a sequence of symbols belonging to a finite and fixes alphabet (Abramson, 1980) whose elements form a data structure. These symbols are chosen with a fixed law of probability and we will admit that they are statistically independent. The probabilities with which the symbols are presented are $p(\Delta_1), p(\Delta_2), ..., p(\Delta_N)$. The quantity of information generated by the occurrence of Δ_i is:

$$I(\Delta_i) = \log \frac{1}{p(\Delta_i)} = -\log p(\Delta_i) \tag{6}$$

It is also called value of surprise of symbol. The formula for the calculation of the mean quantity of information $I(\Delta)$ associated to the source Δ is:

$$I(\Delta) = \sum_{\Delta} p(\Delta_i) \cdot I(\Delta_i) \tag{7}$$

That is to say, they are taking the surprise values of each one of the possibilities of the source Δ and they are pondered according to the occurrence probability $p(\Delta_i)$. The sum of all they will be the quantity of information generated by the source Δ . To measure of $p(\Delta_i)$ approaches to 1, the quantity of information associated with the occurrence of the symbol Δ_i spreads to 0. In the case limit in that the probability of the symbol is 1, the occurrence of Δ_i doesn't generate any information. That is to say, information is not generated by the occurrence of symbols for which alternative possibilities don't exist.

We will designate like Σ to a receiver of information on Δ . To Σ will be denominated sign and it will be formed by the elements of the alphabet A-D, that is to say, the union of set W of characters used to express measurable attributes and to set Φ of n-order monoads that are before the interaction in the " soup-state " (Figure 1).

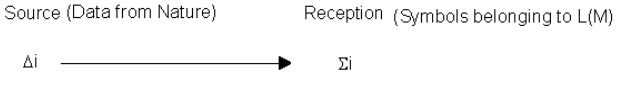

Figure 1. Source-reception process

That way I(Σ) is received information of Δ or about Δ ?. $I_\Delta(\Sigma)$ will be used to designate this new information, indicating the subindex Δ the part of I(Σ) that received information is of Δ . The information transmitted from Δ to Σ is the total quantity of available information in Σ ,

I(Σ), fewer a quantity R or noise, and it will be expressed as:

$$I_{\Delta}(\Sigma) = I(\Sigma) - R \tag{8}$$

In the same way

$$I_{\Delta}(\Sigma) = I(\Delta) - \varepsilon \tag{9}$$

being ε the equivocity of the information generated in Δ which is not transmitted to Σ. The information generated in Δ is divided in two parts:

1. The part $\left[I_{\Delta}(\Sigma)\right]$ that is transmitted to Δ.
2. The part ε that is not transmitted or equivocity. At the same time, the information that there is in Σ can be divided in a similar way in two parts:
   a. The one $\left[I_{\Delta}(\Sigma)\right]$ that represents the received information of Δ.
   b. The remaining part whose source is not Δ or noise R. An increase of R makes to be hidden a part of the sign Σ, and this way $I_{\Delta}(\Sigma)$ will decrease by means of the increase of the equivocity ε.

If the noise increases the quantity of information that gets lost, it diminishes the quantity of transmitted information, but if it doesn't affect to Δ, then $I_{\Delta}(\Sigma)$ continues being the same one.

In the classic Theory of the Information (Abramson, 1980), an equivalence settles down among R and ε, resultant of having chosen the set of possibilities of the source Δ and of the receiver Σ in such a way that $I(\Delta) = I(\Sigma)$. If we imagine changes in the set of possibilities that define $I(\Sigma)$ without the corresponding changes in the set of possibilities that define $I(\Delta)$ and vice versa, have not to exist a necessary equivalence among them. In case there was a maximum dependence among what happens in Δ and what happens in Σ, then R = ε = 0 and the quantity of transmitted information $I_{\Delta}(\Sigma)$ will be highest and in this case $I_{\Delta}(\Sigma) = I(\Sigma)$. Let $p(\Sigma_i / \Delta_i)$ be the conditional probability of Σ_i given Δ_i. One will be able to calculate the contribution of Δ_i to the noise R by means of

$$R = -\sum_{\Delta} p(\Sigma_i / \Delta_i) \cdot \log p(\Sigma_i / \Delta_i) \tag{10}$$

The equivocity ε will be calculated in a similar way

$$\varepsilon = -\sum_{\Delta} p(\Delta_i / \Sigma_i) \cdot \log p(\Delta_i / \Sigma_i) \tag{11}$$

The flow of information depends on underlying causal processes. But it will be necessary to distinguish between the causal relationships and the existent informational among Δ and Σ. If the real data were always the same ones, that is to say there was an experimental perseverance (what doesn't happen in the reality) there would be a strict causal dependence (strong) among Δ and Σ.

It happens that $\{\Delta_i\}$ is cause of $\{\Sigma_i\}$ depending on the real data. The sign Σ_j doesn't say what happened exactly in the source Δ, while $\Sigma_1, \Sigma_2, ..., \Sigma_n$ they say it. From an informational point of view, $\Sigma_1, \Sigma_2, ..., \Sigma_n$ take more information than Σ_j it has more than enough what happened in Δ. Although, for a certain structure of data, each symbol possesses a significance or concrete adjustment, the temporary change of this structure, it can determine a change in the symbol that represents it, a change in its significance or adjustment of the symbol to the fact and in its significant one or decoding on the part of observer (Villacampa et al., 1999a).

3. Text and Indetermination

What one can know that is the relative frequency of every symbol in determined model-text, frequency that differs more or less its probability of occurrence or apparition in the text. However, according to the law of the great numbers, one knows that the observed relative frequency offers toward the probability of apparition, which is said, toward the relative frequency when the total strength or size of the sample increases. The symbol to which belongs the individual apparition is uncertain and the degree of uncertainty or indetermination on the result is function of the relative frequencies of symbol in the text, therefore of its diversity. Consider a lexicon L', $L' \subseteq L$. Suppose a text formed by S complex sentences (state equations) and in their right hand (functions of flow) of each one of them, their simple sentence are formed by SLUNs or CLUNs, generated through the different generative grammars starting from ω variables or different primitive symbols. Each primitive monoad generates an associative field whose number of elements (SLUNs) comes given by $\frac{m^{n+1}-1}{m-1}$ being m the number of first-order monoads (determined by the modeler) and n the order of the monoad of same or superior order to 2. (Villacampa & Usó-Domènech, 1999). The number of possible different SLUNs for complex sentence in a text T write in L' will be

$$Q_{CS} = \omega\left[1 + \frac{m^{n+1}-1}{m-1}\right] \tag{12}$$

The number of possible different SLUNs for text T write in L' will be then:

$$Q_T = S\omega\left[1 + \frac{m^{n+1}-1}{m-1}\right] \tag{13}$$

Let T a text write in L', with N the number total of symbols and Q_T the number of differnt symbols, such as $N \geq Q_T$. If it only exists one symbol in L', $Q_T = N = 1$ there is not any indetermination because the result is certain. If it exists two symbols, one very abundant n_1 in its apparition in all texts and the other very rare n_2, $Q_T = 2 < N, n_1 > n_2, N = n_1 + n_2$, one has the big odds to get the first and the indetermination is weak. The indetermination will be maximal if there are two symbol even having relative abundance of apparition, because in this case, there not are not more of odds to apparition one that the other. If instead of two symbols having the equal relative abundances $n_1 = n_2$ and therefore of the identical

apparition probabilities, the indetermination will be even bigger. It is therefore important to be able to encode the degree of indetermination. One will first consider the case of a text T understanding the same abundant symbols n. The structure of T is compliant to the following diagram:

Present symbol in the text T	$s_1, s_2, s_3, ..., s_{Q_T}$
Probabilities of apparition or relative frequencies $N = \sum n$	$\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}$

The indetermination on the result of an apparition must be an increasing function of N, because as well as it has been said before, more N are big more it is difficult to predict the result of an apparition. Besides, the function must annul himself for N = 1. We will write therefore: Indetermination = f(N) and f(1) = 0. To determine the shape of function f(N), it is necessary to impose a supplementary condition. If in a text two symbols s_1, s_2, appear, with $n_1 = n_2 = n$, the result of an apparition not influencing on the other, the double apparition has n.n possible and also likely compositions. The corresponding indetermination is therefore $f\left(n^2\right)$. One will impose to be as the sum of the indetermination on each of apparitions. It comes back to impose to function f(n) to satisfy to the condition $f\left(n^2\right) = 2f(n)$. One demonstrates that the only function that satisfies to these various conditions at a time is the logarithmic function. One will define the indetermination therefore by the logarithm of the number of possible and also likely cases, either for an apparition: Indetermination = logn. The indetermination will be expressed therefore in bits by the logarithm of basis 2 of possible cases and likely: $In\det er\min ation = \log_2 n$. In one text T as the one considered all symbol play an equivalent role and one can say that each of the present symbols introduced an element or an equal indetermination part to the total indetermination divided by n, either $\frac{1}{n\log_2 n}$. This expression can write himself again $-\frac{1}{n\log_2 \frac{1}{n}}$, this shape having the advantage to make only intervene the relative frequency of symbol $\frac{1}{n}$. The total indetermination in bits $\log_2 n$, will be as the sum of n elements of indetermination introduced by every symbol:

$$In\det er\min ation = -\sum \frac{1}{n}\log_2 \frac{1}{n} \qquad (14)$$

Actually, the words of ours text T doesn't correspond to the considered previously simplistic diagram. In a text understanding Q_T different symbols, each has a relative abundance different of others and the diagram is the next one:

Present symbol in the text T	$s_1, s_2, s_3, ..., s_{Q_T}$
Probabilities of apparition or relative frequencies	$p_1, p_2, p_3, ..., p_{Q_T}$

And

$$In\det er\min ation = -\sum_{i=1}^{Q_T} p_i \log_2 p_i \qquad (15)$$

or Shannon's formula.

The maximal value corresponds to the theoretical case, impossible in ours theory, where all symbol of the text would even have relative abundance n. One recovers: $In\det er\min ation \quad \max imal = \log_2 n$. The minimal value would be gotten in the case of a text T with N symbols where $Q_T - 1$ symbols are represented by only one individual and a symbol by all other individuals. The probability of apparition of the abundant symbol would be very neighbor of 1 and the probability of others apparition very neighbor of zero. The indetermination would be a sum of $Q_T - 1$ terms all neighbors of zero. Finally, to a text of composed diversity given N symbols corresponds a indetermination of which the value expressed in bits is understood between 0 and $\log_2 N$. Here, N is equivalent to text volume Λ and the indetermination is between 0 and $\log_2 \Lambda$.

We consider a text T understanding N individuals distributed between Q_T different symbols having some effectives $n_1, n_2, ..., n_{Q_T}$ like a composed message of different Q_T signals and that brings certain information on the composition and the structure of the source of reality of which has been extracted. There is a deep analogy between the two notions of information and indetermination. It understands himself comfortably if one considers information as the difference between the indetermination on the composition of the message before and after that it is known. If the message has K possible and also likely compositions, the indetermination before is equal to $\log_2 K$. The indetermination a posteriori is hopeless since the message being known, there is not any uncertainty on its composition. The information is $\log_2 K - 0 = \log_2 K$.

In a text of known composition, it is considered the assignment of an individual to some symbol like an elementary signal and the problem is to value the number of compositions possible and also likely of the message. So each individual was identifiable and that one noted the order in which individuals appeared, the text would be assimilated to a composed message of Q_T different signs placed the one following others in a determined order. The number of possible and also likely arrangements that one can get as arranging N objects the some following others is given by N!. Therefore, the number of possible and also likely compositions K will be:

$$K = \frac{N!}{n_1!n_2!...n_{Q_T}!} = \frac{N!}{\prod_{1=1}^{Q_T} n_i!} \qquad (16)$$

The total information in bits will be equal to:

$$\log_2 K = \log_2 N! - \sum_{i=1}^{Q_T} \log_2 n_i! \qquad (17)$$

In the text T, and fixed by their argument (Villacampa et al. 1999a), will come certain the number of state

equations (complex sentences). Therefore, the differential symbol can be considered as not movable. However, the SLUNs form part of the flow equation or sentence where the possibility of exchanges can exist.

If we call $f(s_i), i = 1,...,N$ to the frequencies of each one of the SLUNs or symbol, then the number of possible and equally probable compositions for complex sentence can be calculated using the equation (16)

$$K = \frac{N!}{f(s_1)!\, f(s_2)!... f(s_{Q_T})!} = \frac{N!}{\prod_{1=1}^{Q_T} f(s_i)!} \tag{18}$$

and (17)

$$\log_2 K = \log_2 N! - \sum_{i=1}^{Q_T} \log_2 f(s_i)! \tag{19}$$

When the effectives q_i are all big and in the measure where they are it, one can replace the factorial $f(s_i)!$ by values approached given by the formula of Stirling $f(s)! = \sqrt{2\pi f(s)}\left(\frac{f(s)}{e}\right)^{f(s)}$ .And when strengths are all very big by $f(s)! = \left(\frac{f(s)}{e}\right)^{f(s)}$. One has then:

$$\log_2 K = N \log_2 \frac{N}{e} - \sum_{i=1}^{Q_T} f(s_i) \log_2 \frac{f(s_i)}{e} \tag{20}$$

And as $\sum_{i=1}^{Q_T} f(s_i) = N$ one can write

$$\begin{aligned} \log_2 K &= \sum_{i=1}^{Q_T} f(s_i) \log_2 \frac{N}{e} - \sum_{i=1}^{Q_T} f(s_i) \log_2 \frac{f(s_i)}{e} \\ &= -\sum_{i=1}^{Q_T} f(s_i) \log_2 \frac{f(s_i)}{N} \end{aligned} \tag{21}$$

One doesn't change in anything the value of $\log_2 K$ while multiplying each of terms under the sign $\sum$ by $\frac{N}{N}$ and putting N then in factor. One gets so the expression

$$\log_2 K = -N \sum_{i=1}^{Q_T} \frac{f(s_i)}{N} \log_2 \frac{f(s_i)}{N} \tag{22}$$

Under this shape, one notes that the total information offers, when all effectives q_i are very big, toward a value that only depends of the relative frequencies and the strength on total N. It is therefore logical, since one it is interested to the structure of the text not to consider any information total, $\log_2 K$ but the average information by individual

$$\frac{1}{N} \log_2 K = \frac{1}{N} \log_2 \frac{N!}{\prod_{i=1}^{Q_T} f(s_i)!} \tag{23}$$

This average information expresses in bits and under reserve that all $f(s_i)$ are sufficiently big and offer toward the value $-\sum_{i=1}^{Q_T} \frac{f(s_i)}{N} \log_2 \frac{f(s_i)}{N}$ that is not other that the formula of Shannon:

$$I_{SH} = -\sum_{i=1}^{Q_T} \frac{f(s_i)}{N} \log_2 \frac{f(s_i)}{N} \tag{24}$$

The first term of the equation (23) is similar to

$$\frac{1}{N} \log_2 K = -\sum_{i=1}^{Q_T} \frac{f(q_i)}{N} \log_2 \frac{f(q_i)}{N} \tag{25}$$

and as the second term is the index of diversity of Shannon (24), the formula (25) will be:

$$I_{SH} = \frac{1}{N} \log_2 K \tag{26}$$

after (23) the following index of diversity will appear

$$I_T = \frac{1}{N} \log_2 \frac{N!}{\prod_{i=1}^{Q_T} f(q_i)!} \tag{27}$$

index of diversity that is equivalent to the one obtained by Margalef (1958) for the species of an animal population in an ecosystem.

The average information by individual, gives by the formula (27), offers toward information given by the formula (26) when all strengths of symbol are sufficiently big. In the practice, the two formulas give the same appreciably value when samples are of large size, but values are much more different than samples which total effectives are weaker.

4. Conclusions

The index of diversity is again an evaluation of the diversity since its value offers toward the one of I when N increases, but it is a biased evaluation because values of I_T are always lower to those given by the formula (26). One can convince itself by the following reasoning: To pass of (27) has (26), is sufficient to replace the factorial by their approached values deducted of the formula of Stirling. However this one always drives to approximations by default, of as much less good than he is about the less elevated numbers. In the formula (27), some of numbers that represent to the denominator are rare symbol frequencies and are not in general very elevated. The formula of Stirling conducted in this case to an evaluation by default of the denominator and by excess of the logarithm. In other terms, the formula (26) gives a value in bits superior to the one given by the formula (27) and gaps are of as much bigger than the size of the text is weaker. The bias comes from that, in the calculation of information formulated by (27), one not only knows to the departure the relative different symbols frequencies in the text, but as the strength total N. When one calculates the information signal by signal, the first will have, whatever is the chosen order, the probability of apparition is equal

to relative frequency in the text, and will provide an equal information quantity to the one calculated by the formula (26). But for the following signals, probabilities will change by the fact of the previous signal knowledge. In particular, the last signal will provide hopeless information because it will perfectly be determined. The average information by signal decreases regularly the first at last and the general average will be weaker than the value given by the formula (26).

References

[1] Abramson, N. 1980. *Information Theory and Coding.* McGraw-Hill Book Company, Inc.

[2] Chomsky. N. 1965. *Aspects of the Theory of Syntax*. MIT Press. Cambridge, Massachusets.

[3] Chomsky. N. 1969. *Syntactic Structures*. Mouton. La Haye.

[4] Davies, P. 1983. *God and the new physic.* J.M. Dent & Sons Ltd. London.

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

[6] Gash, H. 2014. Fixed or Probable Ideas. *Foundations of Science.* In press.

[7] Mandelbrot, B. 1954. *Structure formelle des textes et communication. Deux études.* Word, 10. 1-27.

[8] Mandelbrot, B. 1961. *Word frequencies and Markovian models of discourse.* In: Structure of Language and its Mathematical Aspects, Proceedings of Symposia in Applied Mathematics, 12. American Mathematical Society, Providence, Rhode Island, 190-219.

[9] Marashi, S-A.; Tefagh, M. 2014. A mathematical approach to emergent properties of metabolic networks: Partial coupling relations, hyperarcs and flux ratios. *Journal of Theoretical Biology.* 355. pp. 185-193

[10] Margalef, R. 1958. Information theory in Ecology. *International Journal of General Systems.* 3, 36-71.

[11] Nescolarde-Selva, J.A and Usó-Doménech, J.l. 2013. An introduction to Alysidal Algebra (V). *Kybernetes.* Vol. 42 (8), pp. 1248-1264.

[12] Rubin, K. J.; Lawler, K.; Sollich, P.; Ng, T. 2014. Memory effects in biochemical networks as the natural counterpart of extrinsic noise. *Journal of Theoretical Biology*. 357. pp. 245-267.

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

[14] 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.

[15] Usó-Domènech, J.L**.**, Mateu, J. and Lopez, J.A. 2000[a]. MEDEA: software development for prediction of Mediterranean forest degraded areas. *Advances In Engineering Software*. 31, pp 185-196.

[16] Usó-Domènech, J.L**.,** Villacampa, Y., Mateu, J., and Sastre-Vazquez, P. 2000b. Uncertainty and Complementary Principles in Flow Equations of Ecological Models. *Cybernetics and Systems: An International Journal.* 31(2), pp 137-160.

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

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

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

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

[21] Usó-Domènech, J.L, Nescolarde-Selva, J.A. and Lloret-Climent, M. 2014. Behaviours, processes and probabilistic environmental functions in h-open systems. *American Journal of Systems and Software*. Accepted.

[22] Villacampa, Y. and Usó-Domènech, J.L. 1999. *Mathematical Models of Complex Structural systems.* A Linguistic Vision. *International Journal of General Systems*. Vol 28, no1, 37-52.

[23] Villacampa, Y., Usó-Domènech, J.L**.,** Mateu, J. Vives, F. and Sastre, P. 1999[a]. Generative and Recognoscitive Grammars in Ecological Models. *Ecological Modelling*. 117, 315-332.

[24] Villacampa-Esteve, Y., Usó-Domènech, J.L., Castro-Lopez-M, A. and P. Sastre-Vazquez. 1999[b]. *A Text Theory of Ecological Models. Cybernetics and Systems: An International Journal*. Vol 30, 7. 587-607.

[25] Yan,W; Sun, M.; Hu, G.; Zhou, J.; Zhang, W; Chen, J.; Chen, B.; Shen, B. 2014. Amino acid contact energy networks impact protein structure and evolution. *Journal of Theoretical Biology*. 355, pp. 91-104.

[26] Zhu, W.; Hou, J.; Phoebe Y. P. 2010. Semantic and layered protein function prediction from PPI networks. *Journal of Theoretical Biology*. 267. pp. 129-136.

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