Information Approach to Co-occurrence of Words in Written Language

Damián G. Hernández
Consejo Nacional de Investigaciones Científicas y Técnicas
Centro Atómico Bariloche and Instituto Balseiro
8400 Bariloche, Río Negro, Argentina

In this paper we study the distribution of words across the different parts of a book using tools from information theory. In particular, the mutual information between words in the text and parts of the text is compared with the mutual information of a shuffled version of the book. This analysis allows us to extract not only relevant words of the text but also relationships between the different words, such as co-occurrence and repulsion between them. With the connections due to co-occurrence of words, we show how to construct a network that reflects the semantic organization of the book. This method can be applied to other types of sequences, measuring the relations between the different symbols that compose such sequences.

1. Introduction

Understanding the relation between the order of symbols in a complex sequence and how the sequence is formed has been a central question in the study of discrete systems [1–4].

Human language, which ultimately can be expressed in a sequence of words, is a paradigmatic example in this field, as it projects the complexity of the human brain into a one-dimensional sequence. Language has evolved under the pressure of its fundamental function, namely, the exchange of information [5, 6], reaching a state that presents rich organizational structures.

Statistics and information theory approaches have been useful tools in understanding these structures in language. Shannon, in one of his first works on information theory [7], proposed a way of measuring the entropy of printed English through the prediction of letters. The entropy in written language was estimated using compressing codes in [8], while the mutual information between letters was measured in [9], showing the presence of long-range correlations in written texts. More recently, the relationship between the grammatical and semantic structures of written texts and their long-range correlations was analyzed [10, 11].

Complex Systems, **24** © 2015 Complex Systems Publications, Inc.
https://doi.org/10.25088/ComplexSystems.24.2.127

In this paper we study written texts through the mutual information between different parts of the text and the symbols of these sequences, namely, the words. In this sense, we generalize the idea presented by Montemurro and Zanette [12] and apply it not only to individual symbols but to groups of words, which allows us to find quantitative relations between the symbols at different scales. In Section 2 we illustrate the concepts involved through examples. In Section 3 we measure the mutual information between parts and symbols, reviewing some results of previous studies [12]. Section 4 shows how to obtain connections between words using mutual information, while Section 5 introduces analytical generalization to larger groups of symbols. Section 6 shows a preliminary comparison of these measurements between books in Spanish and English, and finally the results are discussed in Section 7.

2. Conceptual Examples

Imagine a sequence of 100 balls, where five of them are red and the rest black. The sequence is divided into 10 parts of equal size. Now we select a part at random, take a ball out of this part, and it happens to be a red ball. The question that arises is, does the fact that the ball is red tell us something about which part we took it from? The answer to this question will depend on how the five red balls are distributed over the 10 parts. For example, if all of them are in the second part, having taken a red ball tells us exactly from which part out of 10 it comes.

Let us consider a new situation where there is a sequence of 100 balls: five of them are red, five are blue, and the rest black. As in the previous case, the sequence is divided into 10 parts. Now we take two balls out of a random part: one is red and the other blue. If the balls were located at random in the sequence, we could calculate the probability of finding a certain number of red and blue balls in a part, as given by a hypergeometric distribution. Then we would expect on average that only two parts will have at least a red ball and a blue ball, so having taken these two balls tells us a great deal about where they come from. However, if in the process of construction of the sequence there was a tendency of the red and blue balls to appear together, we would expect to find them in more parts. So for this last case, taking these two balls tells us less about their origin than in the random case. A similar argument but in the other direction can be made when the balls tend to be apart.

The way to quantify how much the color of the balls tells us about the part of the sequence they come from is the Shannon mutual information between these two variables. Moreover, comparing it with the

https://doi.org/10.25088/ComplexSystems.24.2.127

corresponding information of a shuffled sequence (i.e., taking the difference of information between the sequence and a shuffled sequence without order) allows us to infer how much the mechanism of construction of the sequence differs from a random collocation.

3. Information between Words and Parts of a Text

So we now proceed to analyze books as sequences of words, reviewing some results from [12] where a characteristic length and a list of relevant words are found using this method. As we stated previously, we wish to relate the distribution of each word across the text with the role, grammatical or semantic, that such a word plays.

So consider a text of length N divided into P parts of size $s = N/P$. A word is taken out of a part, and we want to evaluate if that word tells us where it comes from, or the other way around, which words are likely to come from this part. The Shannon mutual information between the parts J and the words W is the difference of the entropy of the parts $H(J)$ minus the entropy of the parts given the words $H(J \mid W)$ [13], and it can be expressed as

$$I(J, W) = H(J) - H(J \mid W) = \sum_{w=1}^{K} p(w) \sum_{j=1}^{P} p(j \mid w) \log_2\left(\frac{p(j \mid w)}{p(j)}\right), \tag{1}$$

where the index w runs over the words (considering a vocabulary of K words) and the index j runs over the parts of the text. As all the parts have the same size $N_j = s$, it implies that the marginal probability of a part is $p(j) = s/N = 1/P$. To calculate the conditional probability $p(j \mid w)$, that is, the probability of being in part j given the word w, we use the Bayes rule,

$$p(j \mid w) = \frac{p(w \mid j)\, p(j)}{p(w)}, \tag{2}$$

where the probability of taking the word w given that we are in part j is $p(w \mid j) = n_w^{(j)}/s$, with $n_w^{(j)}$ being the number of times the word w appears in the part j. The normalization factor corresponds to

$$p(w) = \sum_{i=1}^{P} p(w \mid i)\, p(i) = \frac{n_w}{N}, \tag{3}$$

that is, the frequency of the word in the whole text. So finally we have

$$p(j \mid w) = \frac{n_w^{(j)}}{n_w}, \tag{4}$$

https://doi.org/10.25088/ComplexSystems.24.2.127

and we are able to calculate the mutual information between each word and the parts of the text for a scale s.

As we like to observe how the construction of the sequence differs from a random shuffle of its symbols, we subtract from this information the information corresonding to a shuffled text $\langle \hat{I}(J, W) \rangle$, where the average is taken over all possible shuffles. So by measuring the difference $\Delta I_1(s) = I(J, W) - \langle \hat{I}(J, W) \rangle$, we are taking as reference a shuffled version of the text where there is still information, due to expected fluctuations in the distribution of words.

The quantity $\Delta I_1(s)$ splits naturally into the contributions of the different words as $\Delta I_1(s) = \sum_w \Delta I_{\{w\}}(s)$. Each term, for a specific scale and word, can be positive if the word presents a larger heterogeneity than in a shuffled text, or it can be negative if it has a larger homogeneity. We have to take into account that each term $\Delta I_{\{w\}}(s)$ is weighted by $p(w)$ (i.e., the frequency of the word), so that the interplay between the frequency and the heterogeneity will determine the contribution of the corresponding term. Another possible form for this expression, considering the first line of equation (1), is

$$\Delta I_1(s) = \sum_{w=1}^{K} p(w) \left(\langle \hat{H}(J \mid w) \rangle - H(J \mid w) \right), \tag{5}$$

where the entropy of the parts for a given word w is

$$H(J \mid w) = -\sum_{j=1}^{P} p(j \mid w) \log_2(p(j \mid w)). \tag{6}$$

The calculus for the entropy $\langle \hat{H}(J \mid w) \rangle$ of the shuffled text is provided in Appendix A.

It is important to highlight that this measure possesses some symmetries; that is, its value remains the same if we make some changes to the text. As it only uses the occurrences of words in each part, the information does not vary if we change the order of words inside a part nor if we swap parts. In this sense, we believe that this approach is the very next step after analyzing word frequency in the whole text.

Figure 1 shows the difference of mutual information for three books as a function of the scale s. The texts are *The Prince and the Pauper* by Mark Twain, *On the Origin of Species* by Charles Darwin, and *The Analysis of Mind* by Bertrand Russell. The curves are similar in the three cases, presenting a maximum around $s \simeq 1000$ (scale related to the semantic structure of the text) and they become negative around $s \simeq 50$. The maximum at $s \simeq 10^3$ is explained by assuming that there are words whose distributions are concentrated and bounded in blocks of length $\sim 10^3$. So the meaning assigned to this

https://doi.org/10.25088/ComplexSystems.24.2.127

scale corresponds to the length in words in which the author of the book tends to write about the same subject on average [12].

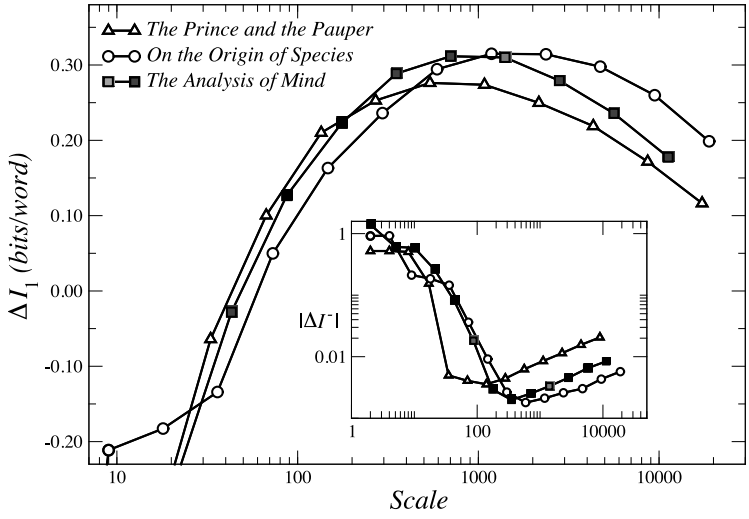

Figure 1. Difference of mutual information for three books as a function of the scale s. The texts are *The Prince and the Pauper* by Mark Twain, *On the Origin of Species* by Charles Darwin, and *The Analysis of Mind* by Bertrand Russell. The inset shows a log–log plot of the absolute value of the negative component of $\Delta I_1(s)$.

At the second scale, $s \simeq 50$, a random word taken from the text is not going to give us information about which part it belongs to; that is, on average the words have the same heterogeneity as a random shuffle of the text. However, a specific word may contribute with a negative term (i.e., a loss of information $\Delta I_{\{w\}}(s) < 0$), due to a homogeneous distribution through the book at that scale, while another word may do it with a positive term, $\Delta I_{\{w\}}(s) > 0$. The inset of Figure 1 shows in a log–log plot the absolute value of the negative contribution of $\Delta I_1(s)$, which is highly dominated by the most frequent words at shorter scales.

At the scale where the information between words and parts reaches a maximum, we can make a list of the words ordered by their contribution $\Delta I_{\{w\}}(s)$. These words are the ones that possess a combination of high frequency and heterogeneity.

Table 1 shows the first 15 words that contribute the most at the scale of the maximum of $\Delta I_1(s)$ for the three books previously mentioned.

https://doi.org/10.25088/ComplexSystems.24.2.127

Figure 2 shows the occurrence of the words *hybrids* and *varieties* through the parts of the book *On the Origin of Species* for a scale $s = 1182$. As can be observed, these words have a large heterogeneity throughout the text, as expected.

The Prince and the Pauper	*On the Origin of Species*	*The Analysis of Mind*
i	species	image
she	varieties	images
her	hybrids	belief
he	forms	word
the	islands	memory
tom	selection	words
of	genera	you
prince	will	desire
thou	breeds	sensations
thy	characters	we
my	groups	object
is	seeds	knowledge
me	pollen	a
you	sterility	i
hendon	plants	the
…	…	…

Table 1. Informative words at the maximum of ΔI_1 (i.e., words with highest values of $\Delta I_{\{w\}}(s)$).

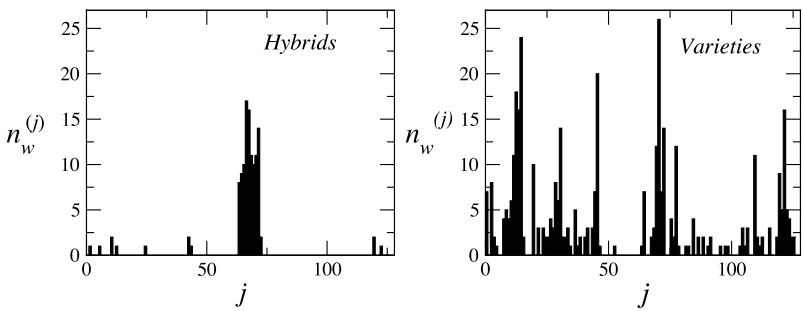

Figure 2. Number of occurrences $n_w^{(j)}$ as a function of the parts j of size $s = 1182$ of the book *On the Origin of Species* for the words *hybrids* and *varieties*.

We observe that there are words with high frequency in the lists, especially for the first book, such as *i*, *will*, *you*; but there also are

https://doi.org/10.25088/ComplexSystems.24.2.127

words with a high semantic content, such as *prince*, *selection*, *memory*, and so forth.

On the other hand, there are some words, especially at shorter scales, that contribute negatively, that is, that have a more homogeneous distribution than they do in shuffled text. As we anticipated, these correspond to the words with higher frequency, such as *the*, *and*, *in*, and others like *but*. This last word possesses a functional use introducing a phrase or clause, contrasting with what has already been mentioned, so we expect to find it many times, but we do not expect—due to its function—to find instances of this word very close to each other. At larger scales, although the negative component is minimal, it is important to observe that the words that do contribute are of a special kind, in general being conjunctions and adverbs such as *however*, *then*, *whereas*, *naturally*, *commonly*, and so forth.

4. Information between Pairs of Words and Parts of a Text

Instead of taking one word of a part of the text, we can take two words and ask the same questions as before, how much information do these words give us about the part of the text we are in? Or conversely, how much information does a specific part of the text give us about the pair of words that came out of it? These statements must always be considered in relation to the shuffled version of the text. We shall see that by doing this, it is possible to find interesting connections between words that possess distributions through the text that are related in some specific way.

The analytical definitions are very similar to the previous case. The difference of the mutual information corresponds to

$$\Delta I_2(s) = \sum_{v,w=1}^{K} \Delta I_{\{v,w\}}(s) = \sum_{v,w=1}^{K} p(\{v, w\}) \left[\left\langle \hat{H}(J \mid \{v, w\})\right\rangle - H(J \mid \{v, w\})\right], \tag{7}$$

where the pair $\{v, w\}$ identifies the two words. The entropy of the parts for a given pair $\{v, w\}$ is

$$H(J \mid \{v, w\}) = -\sum_{j=1}^{P} p(j \mid \{v, w\}) \log_2(p(j \mid \{v, w\})), \tag{8}$$

where in the same way as in the previous section, we can use the Bayes rule to compute $p(j \mid \{v, w\})$, that is, the probability of the part j

https://doi.org/10.25088/ComplexSystems.24.2.127

given the pair $\{v, w\}$,

$$p(j \mid \{v, w\}) = \frac{p(\{v, w\} \mid j)\, p(j)}{\sum_i p(\{v, w\} \mid i)\, p(i)}. \tag{9}$$

In order to calculate the probability of extracting the pair $\{v, w\}$ out of the part j, we need to consider if we are taking the words with or without replacement, although the difference is minimal. In this paper, we choose to do it without replacement, so that

$$p(\{v, w\} \mid j) = \begin{cases} \dfrac{2\, n_v^{(j)}\, n_w^{(j)}}{s\,(s-1)} & \text{if } v \neq w \\ \dfrac{n_v^{(j)}\,\left(n_v^{(j)} - 1\right)}{s\,(s-1)} & \text{otherwise.} \end{cases} \tag{10}$$

In the case when the words are different, combining equations (9) and (10) we obtain

$$p(j \mid \{v, w\}) = \frac{n_v^{(j)}\, n_w^{(j)}}{\sum_{i=1}^{P} n_v^{(i)}\, n_w^{(i)}}. \tag{11}$$

We observe that the probability $p(j \mid \{v, w\})$ used in the entropy for this case is proportional to the product of the occurrences of the words, so it will be different from zero only when both words appear in the part j. This means, considering equations (7) and (8), that if the words are homogeneously distributed in m parts in which they both appear, the entropy of the parts given the pair will be $H(J \mid \{v, w\}) \sim \log_2(m)$. So basically, $\Delta I_{\{v,w\}}(s)$ is measuring if the words concur in more or fewer parts than in a shuffled text, and weighting it with the frequency of the pair $\{v, w\}$. The entropy for the shuffled text $\left\langle \hat{H}(J \mid \{v, w\}) \right\rangle$ is calculated in a similar way as before (see Appendix A). The marginal probability of the pair, if we take the words without replacement, is

$$p(\{v, w\}) = \frac{2}{N\,(s-1)} \sum_{i=1}^{P} n_v^{(i)}\, n_w^{(i)}. \tag{12}$$

Here we considered that the words in the pair are different (we checked that the component of ΔI_2 for pairs with the same word repeated represents approximately 0.4% of the total, so we are ignoring it). Evidently now we have an arduous calculation, as ΔI_2 possesses many more terms, just about K^2 (though many of them will be zero).

Figure 3 shows the information encoded between pairs of words and parts of the text ΔI_2 as a function of the scale for the three books previously mentioned. We notice that the curves are similar to those

https://doi.org/10.25088/ComplexSystems.24.2.127

of $\Delta I_1(s)$, as they have a maximum around $s \simeq 1000$ and become negative around $s \simeq 50$.

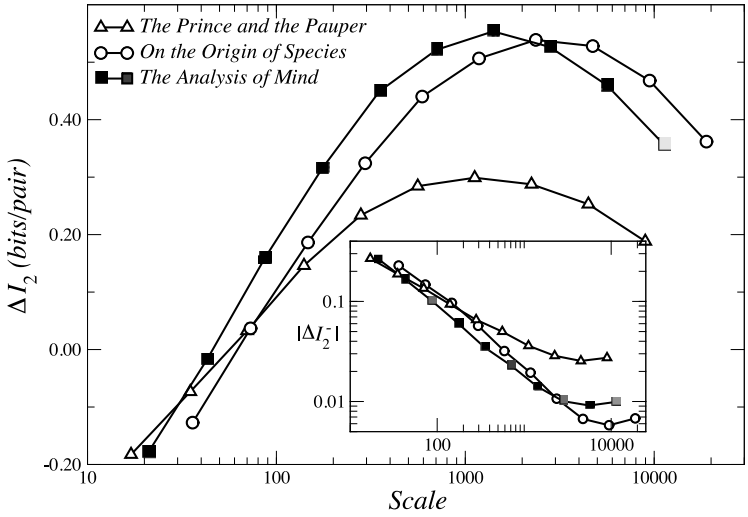

Figure 3. Difference of mutual information, taking two words as a function of the scale s. The insert shows a log–log plot of the absolute value of the negative component of $\Delta I_2(s)$.

At this point we have to draw attention to that similarity in an effort to understand the reason this happens, so let us consider the following facts: in every book there are words that possess a very high frequency (*the*, *of*, *and*, *to*, *a*, etc.) and correspond to an appreciable portion of the total words in a book [14]. Let us use for these words the notation u_i. In most of the scales, these words have a homogeneous distribution through the text, which means that $n_u^{(j)} \approx n_u P^{-1}$. So if we take a pair containing one of these words, the entropy for a pair $\{u, w\}$ is

$$\begin{aligned} H(J \mid \{u, w\}) &= -\sum_{j=1}^{P} p(j \mid \{u, w\}) \log_2(p(j \mid \{u, w\})) \simeq \\ &-\sum_{j=1}^{P} \left(\frac{n_u P^{-1} n_w^{(j)}}{\sum_i n_u P^{-1} n_w^{(i)}} \right) \log_2 \left(\frac{n_u P^{-1} n_w^{(j)}}{\sum_i n_u P^{-1} n_w^{(i)}} \right) = \\ &-\sum_{j=1}^{P} \left(\frac{n_w^{(j)}}{n_w} \right) \log_2 \left(\frac{n_w^{(j)}}{n_w} \right) = H(J \mid w), \end{aligned} \tag{13}$$

assuming that the correlation between these words is negligible. In the

https://doi.org/10.25088/ComplexSystems.24.2.127

same way, it can be shown that $\left\langle \hat{H}\left(J \mid \{u, w\}\right)\right\rangle \simeq \left\langle \hat{H}\left(J \mid w\right)\right\rangle$ with the condition that $n_u \gg P$ (see Appendix A). So considering a pair with a high-frequency word u and another w, its contribution to the information is

$$\begin{aligned} &\Delta I_{\{u,w\}}(s) = p(\{u, w\}) \left[\left\langle \hat{H}\left(J \mid w\right)\right\rangle - H\left(J \mid w\right)\right] = \\ &2 \frac{n_u}{N} \frac{n_w}{N} \left[\left\langle \hat{H}\left(J \mid w\right)\right\rangle - H\left(J \mid w\right)\right] = 2 \frac{n_u}{N} \Delta I_{\{w\}}(s). \end{aligned} \tag{14}$$

This implies that when summing over the pairs $\{u, w\}$, all these terms will contribute to $\Delta I_2(s)$ with an important component that is proportional to $\Delta I_1(s)$. This is the reason we observe in the curves of Figure 3 a similar behavior to that of the curves from Figure 1.

However, a difference can be noted in the negative contribution, and it is that for $\Delta I_2^-(s)$ there is a power-law behavior for $s \in \left(10, 10^3\right)$, while for $\Delta I_1^-(s)$ there is a faster increase as the scale becomes smaller, but it is dominated by a few words with high frequency. This difference is pointing out that a new phenomenon may be occurring for $\Delta I_2^-(s)$ in this scale range.

So in order to analyze what part of $\Delta I_2(s)$ comes from links with a high-frequency word and an informative word from $\Delta I_1(s)$, we consider the first 500 links ranked by their contribution to $\Delta I_2^{\pm}$ and check if each of them is composed of one of the five most frequent words and one of the first 100 words from $\Delta I_1^{\pm}$. Figure 4 shows the fraction of links that fulfill this condition as a function of the scale for

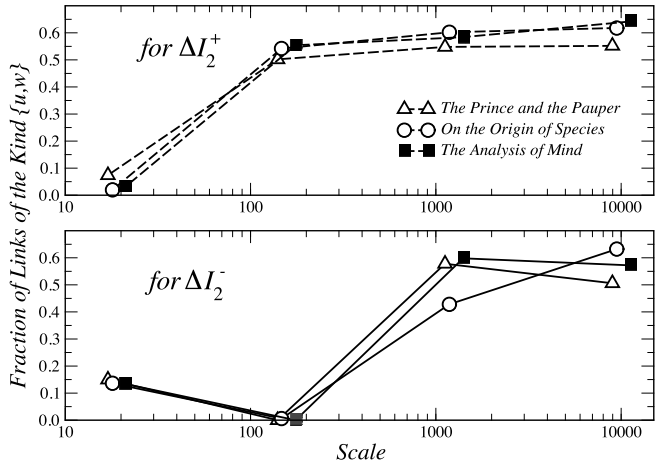

Figure 4. Fraction of links of $\Delta I_2(s)$ composed of a high-frequency word and a word from $\Delta I_1(s)$ as a function of the scale s. We consider the first 500 links ranked by their contribution to $\Delta I_2^{\pm}$, the first 100 words from $\Delta I_1^{\pm}$, and the five most frequent words.

https://doi.org/10.25088/ComplexSystems.24.2.127

both the negative (lower panel) and positive (upper panel) components of ΔI_2. We observe a consistent pattern in the three books analyzed, in which most of the behavior of ΔI_2 for large scales is explained by this phenomenon. In this sense, we can say that such behavior is inherited from ΔI_1.

We observe that this explanation fails for the positive component of ΔI_2 when $s \sim 10$ (just 4% of the links), and for the negative component when $s \sim 100$ (only 0.2% of the links). So to understand how these links are formed, Table 2 shows the links from $\Delta I_2^+ (s = 18)$ and from $\Delta I_2^- (s = 147)$ for the book *On the Origin of Species*.

Links from ΔI_2^+ ($s = 18$)	Links from ΔI_2^- ($s = 147$)
the - a	selection - natural
of - as	organic - beings
the - as	pollen - flower
of - to	and - but
and - or	water - fresh
in - by	life - conditions
a - are	closely - allied
in - as	this - but
of - it	america - south
in - on	bees - wax
in - from	cells - cell
the - their	islands - oceanic
have - be	pigeon - rock
a - as	hybrids - sterility
to - as	bee - hive
…	…

Table 2. Links from ΔI_2^+ ($s = 18$) and from ΔI_2^- ($s = 147$) for the book *On the Origin of Species*.

Considering first the links from $\Delta I_2^+ (s = 18)$, we observe that they are in general two high-frequency words, but with the particularity that they have a slight negative correlation between them. That is, they correspond to prepositions, articles, and conjunctions whose functions somehow overlap, so they tend to be apart from each other at this scale.

In order to see this effect, Figure 5 shows the number of occurrences $n_w^{(j)}$ as a function of j for the first 100 parts of size $s = 18$ of the book *On the Origin of Species* for the pairs {*and*, *or*} and {*in*, *by*}. The fact that these words repel each other implies that

https://doi.org/10.25088/ComplexSystems.24.2.127

$p(j \mid \{v, w\}) \propto n_v^{(j)} n_w^{(j)}$ will be zero in more parts than in a random shuffle of the text (consider that for this scale, $n_w^{(j)} \sim 1$ even for high-frequency words), which results in a gain of information in relation to the shuffled version. This effect of repulsion between words is a new phenomenon that is not inherited from ΔI_1.

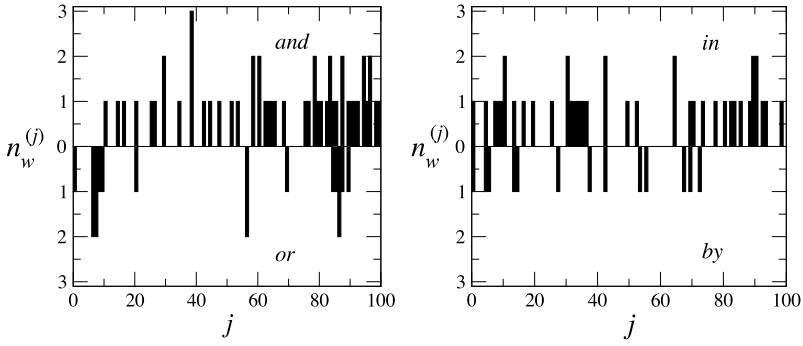

Figure 5. Number of occurrences $n_w^{(j)}$ as a function of j for the first 100 parts of size $s = 18$ of the book *On the Origin of Species* for the pairs {*and*, *or*} and {*in*, *by*} (the occurrences of the second word of the pair is plotted as negative for a better view).

As we previously stated, we also need to explain the negative component of ΔI_2 when $s \sim 100$. We observe that most of the links from $\Delta I_2^- (s = 147)$ are composed of words that are semantically connected. Some of them are linked because they are used together, and others because they share the same context, like {*pollen*, *flower*} and {*males*, *females*}.

Figure 6 shows $n_w^{(j)}$ as a function of j for the parts of size $s = 147$ of the book *On the Origin of Species* for the words {*pollen*, *flower*} and for the words {*water*, *fresh*}. In this scale, the text is divided into $P \sim 10^3$ parts, so these words, which have $n_w \lesssim 10^2$, are absent from many of the parts, but when they are present they tend to appear together in most of the cases. In the same way as the example of the red and blue balls given in Section 2, these words appear together in more parts than they would in a shuffled text. So if the pair is found in more parts, this is equivalent to possessing less information, because it is necessary to ask more questions to infer which part it comes from (i.e., having taken this pair tells us less about where it comes from than it would in a shuffled text). This is the reason these links possess negative ΔI_2 $(H(J \mid \{v, w\}) > \langle \hat{H}(J \mid \{v, w\}) \rangle)$.

It is important to highlight that to have this type of negative link is not only a necessary co-occurrence, but also that the words must be

https://doi.org/10.25088/ComplexSystems.24.2.127

sparse through the different parts. In the books analyzed, this happens at a scale $s \sim 100$, which is close to the scale where the mutual information between parts and pairs of words vanishes (i.e., the information gained by the heterogeneity of the pairs equals the information lost by this co-occurrence of words).

Table 3 shows the links of ΔI_2^- for *The Analysis of Mind* ($s = 175$) and for *The Prince and the Pauper* ($s = 140$).

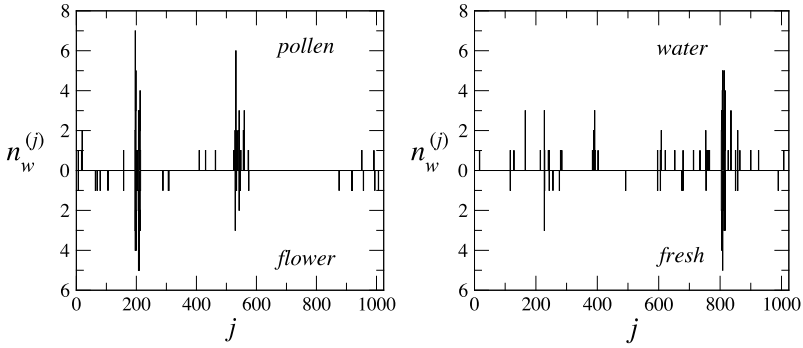

Figure 6. Number of occurrences $n_w^{(j)}$ as a function of j for the parts of size $s = 147$ of the book *On the Origin of Species* for the pairs {*pollen*, *flower*} and {*water*, *fresh*} (the occurrences of the second word of the pair is plotted as negative for a better view).

The Analysis of Mind	*The Prince and the Pauper*
door - window	hendon - miles
more - than	had - been
plato - socrates	prince - wales
left - window	court - offal
door - left	at - last
truth - falsehood	st - john
2 - 1	thou - art
discomfort - pleasure	your - majesty
response - accuracy	thou - lt
appearances - medium	hendon - hall
mnemic - causation	canty - john
appearances - appearance	nan - bet
box - toys	more - than
colour - patch	more - once
self - evidence	your - please
…	…

Table 3. Links from ΔI_2^- ($s = 175, 140$) for the books *The Analysis of Mind* and *The Prince and the Pauper*.

https://doi.org/10.25088/ComplexSystems.24.2.127

4.1 Network of Words

From the list of links ordered by their contribution to ΔI_2^-, such as the ones in Table 3, a network or graph of words can be constructed by progressively adding links. We consider through such a procedure the network for the book *The Analysis of Mind* at the scale $s = 175$ and present some preliminary results about the structure of this kind of network.

Figure 7 shows the number of nodes that belong to the largest component as a function of the links added progressively in accordance with their contribution to ΔI_2^-. A percolation threshold can be observed near 600 links that corresponds to the coalition of some communities to form a giant component. In the inset, the clustering coefficient $C(k)$ for each node is plotted as a function of its degree k, once 10^4 links have been added. Although there is a decrease in the clustering as the degree grows, indicating that highly linked words do not possess interconnected neighbors, there is no clear scaling behavior such as $C(k) \sim k^{-1}$ to ensure that this network can be considered as a hierarchical one [15]. The giant component of the network, after 800 links have been added (i.e., after the percolation threshold), is observed at the right of Figure 7. It shows a clear tree structure with few cycles, which is a positive aspect when trying to classify the different words in communities, as there are few nodes that are difficult to classify.

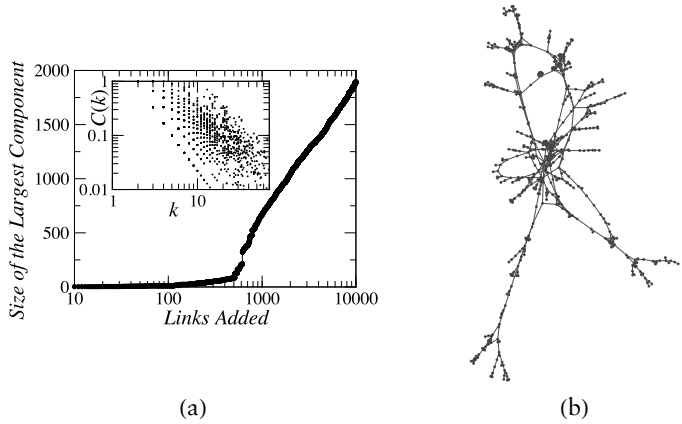

Figure 7. Size of the largest component of the network as a function of the links added progressively in accordance with their contribution to ΔI_2^- (left side). The inset shows the clustering coefficient $C(k)$ for each node as a function of its degree k, once 10^4 links have been added. The network on the right side corresponds to the giant component after 800 links have been added. The book analyzed is *The Analysis of Mind*.

https://doi.org/10.25088/ComplexSystems.24.2.127

Figure 8 shows the second-largest component of the network once 350 links have been added (i.e., before the percolation threshold). On the left side, words mainly related to *desire* are observed, while on the right side are words semantically linked to *beliefs* and *truth*. A clique, a subset in which every two nodes are connected, composed of the words *truth*, *falsehood*, *true*, and *false*, is located in the center of the right side of the graph.

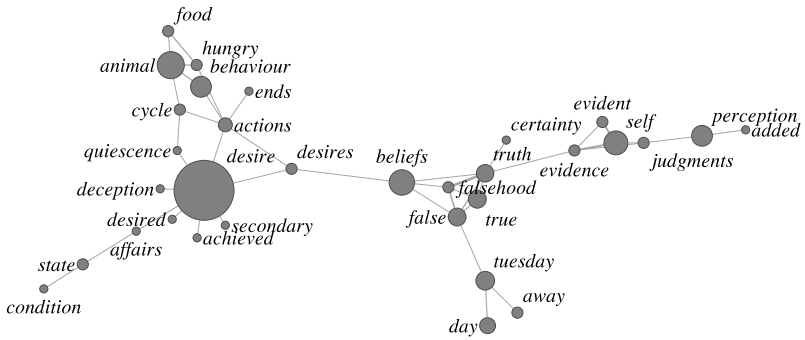

Figure 8. The second-largest component of the network for the book *The Analysis of Mind* once 350 links have been added. Links are obtained from the pair of words that contribute the most to $\Delta I_{\overline{2}}\,(s \sim 10^2)$. The size of the nodes is proportional to $\Delta I_{\{w\}}\,(s \sim 10^3)$, while the thickness of the edges is proportional to $\Delta I_{\overline{\{v,w\}}}\,(s \sim 10^2)$.

5. Generalization to Groups of Words

At this point, we are able to express the analytic generalization to groups of words. In the same way as in the previous section, we consider the mutual information $\Delta I_m\,(s)$ between the parts of a text and m words taken from one of the parts in relation to a shuffled text. In this way, we can find relations or interactions between these groups of m words or m-plets.

Note that in this process the m-plets that we find will inherit the interactions from smaller groups (i.e., $(m-1)$-plets, $(m-2)$-plets, etc.) in the combination with frequent words.

The analytical formulas for ΔI_m are very similar to those of ΔI_2. The probability of each part given the m-plet $\{w_1, w_2, \ldots, w_m\}$ is

$$p\,(j \,|\, \{w_1, w_2, \ldots, w_m\}) = \frac{n^{(j)}_{w_1}\, n^{(j)}_{w_2} \ldots n^{(j)}_{w_m}}{\sum_{i=1}^{P} n^{(i)}_{w_1}\, n^{(i)}_{w_2} \ldots n^{(i)}_{w_m}}, \tag{15}$$

https://doi.org/10.25088/ComplexSystems.24.2.127

while the marginal probability for the m-plet is

$$p\left(\{w_1, w_2, \ldots, w_m\}\right) = \frac{m! \sum_{i=1}^{P} n_{w_1}^{(i)} n_{w_2}^{(i)} \ldots n_{w_m}^{(i)}}{N\left(s-1\right)\left(s-2\right)\ldots\left(s-m+1\right)}. \tag{16}$$

As before, these equations stand when the words in the m-plet are different from each other, although slight modifications are needed when there are words that are repeated. Finally, the mutual information corresponds to

$$\begin{aligned} \Delta I_m(s) = & \sum_{w_1,\ldots,w_m=1}^{K} \Delta I_{\{w_1,w_2,\ldots,w_m\}}(s) = \\ & \sum_{w_1,\ldots,w_m=1}^{K} p(\{w_1, w_2, \ldots, w_m\}) \Big[\\ & \left\langle \hat{H}\left(J \mid \{w_1, w_2, \ldots, w_m\}\right)\right\rangle - \\ & H(J \mid \{w_1, w_2, \ldots, w_m\})\Big]. \end{aligned} \tag{17}$$

Evidently the calculus of ΔI_m is a very arduous task compared to the one from ΔI_2, so in these cases it may be convenient to consider a reduced set of words instead of the whole vocabulary.

6. Comparison between Languages

In this section we are going to present some preliminary results regarding the comparison of some of the previous measurements between two groups of books, one in Spanish and the other in English. Each group contains 100 books, and they have been extracted from the web of Project Gutenberg [16]. We chose books whose plain text size lies within 200 Kb and 600 Kb, so that the lengths of the books are of the same order of magnitude ($N \sim 6\times10^4$).

Figure 9 shows the scale s_1 at which the information ΔI_1 reaches its maximum, and the scale s_2 at which ΔI_2 vanishes, for the books in Spanish and English. The scale s_1 stands for the characteristic length in which the author writes about the same subject, while for the scale s_2 the information lost due to co-occurrences of pairs of words equals the information gained due to heterogeneity in the pairs. Although the clouds of points are mixed, Spanish books present on average larger values for both scales s_1 and s_2.

Figure 10 shows the maximum information per word $\Delta I_1\,(s_1)$ and the negative component of the information between parts and pairs of words $|\Delta I_2^-\,(s_2)|$ at the scale where ΔI_2 vanishes, for the books in Spanish and English. It is clearly observed that the information lost

https://doi.org/10.25088/ComplexSystems.24.2.127

due to co-occurrence at s_2 (i.e., $|\Delta I_2^-(s_2)|$) for most of the Spanish books is lower than for the English ones. In the same way, Spanish books present on average slightly less information per word at the maximum. So the Spanish language carries less information per symbol, and also less information within the interaction of words.

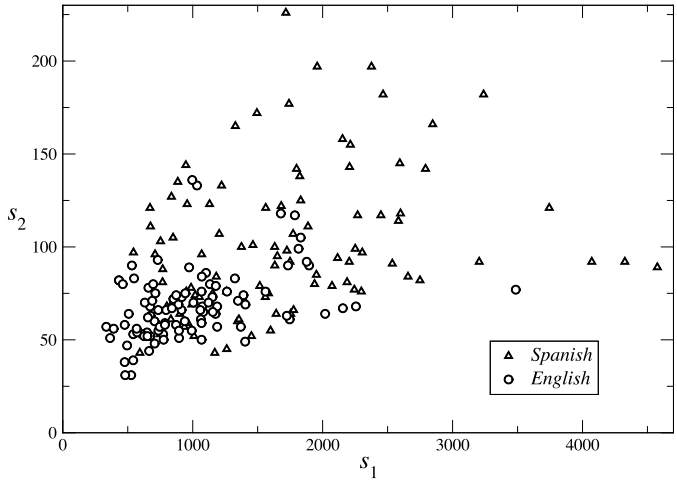

Figure 9. Scale s_1 at which the information ΔI_1 reaches its maximum, and scale s_2 at which ΔI_2 vanishes, for 100 books in Spanish and 100 books in English extracted from Project Gutenberg.

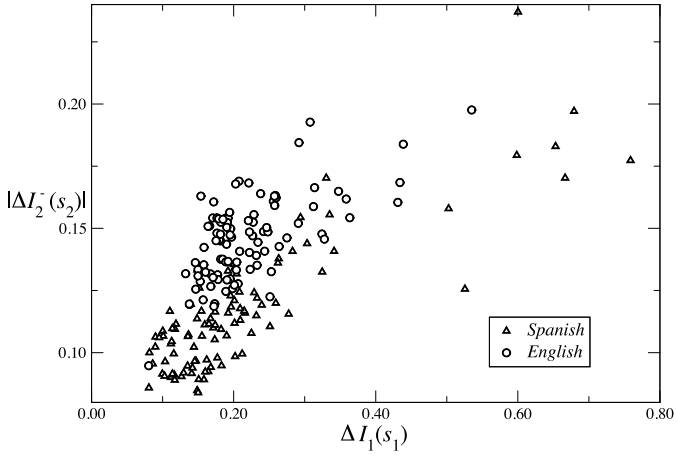

Figure 10. Maximum information per word $\Delta I_1(s_1)$ and the negative component of the information between parts and pairs of words $|\Delta I_2^-(s_2)|$ at the scale where ΔI_2 vanishes, for books in Spanish and English extracted from Project Gutenberg.

https://doi.org/10.25088/ComplexSystems.24.2.127

The combination of possessing larger scales and less information per symbol in Spanish implies that more words are needed to express concepts and ideas than in English. This fact is in agreement with the known phenomenon of "word growth" when translating from English to Spanish, which results in the use of $\sim 25\,\%$ more words.

7. Discussion

In this paper we present a method to analyze finite sequences of words (i.e., books) and find relations between the words based on their distributions throughout the sequences. The method relies on measuring the Shannon mutual information between parts of the texts and the words, in relation to a shuffled version of the texts.

In addition to finding a characteristic scale and relevant words, this method allows us, through the evaluation of the mutual information between parts and pairs of words, to extract different types of interactions between words. At a scale of 20 words, a weak repulsion is found between some frequent words due to their having similar functions, and therefore the probability of appearing together at this scale is less than in the shuffled text. On the other hand, connections between words that co-occur in a sparse way have been found at a scale of 150 words. These interactions happen not only with words that are used together, but also with those that possess a strong semantic link. From this last type of connection, we consider an example of the construction of a network of words for a book. This network presents nearly a tree structure, a fact that allows a good classification of words in different semantic communities.

An analytical generalization for the method is presented, which allows us to account for interactions of groups with a larger number of symbols. Also we compare for two groups of books, one in Spanish and the other in English, some of the quantities defined, showing in particular that the information encoded in words and in pairs is on average larger for the English group.

An unexplored path that may prove valuable to analyze is considering the order of words within a part, since in this case the connections found may be interpreted as causation links, and the corresponding networks, which would become ordered graphs, may bring new insights into the formation of the sequence.

Further studies about the topology of networks of words are an interesting projection of this work. Another aspect to consider in future research is the application of this method to other types of sequences. Even time sequences of events, which can be somehow categorized, are a rich field to test these ideas.

https://doi.org/10.25088/ComplexSystems.24.2.127

Acknowledgments

Helpful and fruitful discussion with Damián H. Zanette is gratefully acknowledged.

Appendix

A. Entropies for the Shuffled Text

The shuffle entropy of the parts given a word $\langle \hat{H}(J \mid w) \rangle$ is the average entropy for a mixed sequence over all possible mixes. An analytical formula is possible to obtain in this case [12]. Recalling equations (4) and (6), we can express the entropy for a mix as

$$\hat{H}(J \mid w) = -\sum_{j=1}^{P} \frac{m_j}{n_w} \log_2\left(\frac{m_j}{n_w}\right), \tag{A.1}$$

where m_j is the number of times the word w appears in part j for the mix and n_w is the total frequency. Taking the average over all possible shuffles,

$$\langle \hat{H}(J \mid w) \rangle = -\sum_{m_1+\cdots+m_P=n_w,\, m_j \le N/P} p(m_1, \ldots, m_P) \sum_{j=1}^{P} \frac{m_j}{n_w} \log_2\left(\frac{m_j}{n_w}\right). \tag{A.2}$$

Marginalizing in each term of the interior sum, this previous equation reduces to

$$\langle \hat{H}(J \mid w) \rangle = -P \sum_{m=1}^{\min\{n_w, N/P\}} p(m) \frac{m}{n_w} \log_2\left(\frac{m}{n_w}\right), \tag{A.3}$$

where $p(m)$ is the marginal probability of finding m instances of the word w in a part and $N/P - m$ instances of words that are not w,

$$p(m) = \frac{\binom{n_w}{m}\binom{N-n_w}{N/P-m}}{\binom{N}{N/P}}. \tag{A.4}$$

If the size of the text N and the number of parts P are fixed, $\langle \hat{H}(J \mid w) \rangle$ is a function only of n_w. For $n_w \gg P$, the words distribute homogeneously through the parts so that $\langle \hat{p}(j \mid w) \rangle \approx 1/P$ and the entropy is

$$\langle \hat{H}(J \mid w) \rangle \approx \log_2(P). \tag{A.5}$$

https://doi.org/10.25088/ComplexSystems.24.2.127

On the other hand, for $n_w \ll P$ only a few parts have one symbol, so that for those parts $\left\langle \hat{p}\,(j \mid w) \right\rangle = 1 \,/\, n_w$ and

$$\left\langle \hat{H}\,(J \mid w) \right\rangle = \log_2(n_w). \tag{A.6}$$

The analytical formula for the entropy of the parts given two words $\left\langle \hat{H}\,(J \mid \{v, w\}) \right\rangle$ is much more complicated, as it is not possible to do the marginalization. Recalling equations (8) and (11),

$$\hat{H}(J \mid \{v, w\}) = -\sum_{j=1}^{P} \frac{m_v^{(j)}\, m_w^{(j)}}{\sum_i m_v^{(i)}\, m_w^{(i)}} \log_2 \left[\frac{m_v^{(j)}\, m_w^{(j)}}{\sum_i m_v^{(i)}\, m_w^{(i)}} \right], \tag{A.7}$$

where $m_v^{(j)}$ and $m_w^{(j)}$ are the frequencies of the words in the different parts for the mixed text.

The analytical formula is impractical, as it involves the joint probability $p\left(m_v^{(1)}, \ldots, m_v^{(P)}, m_w^{(1)}, \ldots, m_w^{(P)}\right)$. However, the average entropy can be estimated by performing shuffles of a sequence composed by n_v symbols of a kind, n_w of another, and $(N - n_v - n_w)$ of a third kind. This estimation can be simplified by considering the sizes of the parts up to some value (e.g., $s_0 \sim 20$), because if sizes beyond that value are needed, it would mean that $n_v \gg P$, and the distribution for such a word can be considered as uniform through the parts, as the instances of this word are randomly distributed over the parts. For these cases, we can consider that $m_v^{(j)} \simeq n_v\, P^{-1}$ and follow the same reasoning as in equation (13), so that

$$\left\langle \hat{H}\,(J \mid \{v, w\}) \right\rangle = \left\langle \hat{H}\,(J \mid w) \right\rangle. \tag{A.8}$$

In this way we just consider $(s_0\, P - n_v - n_w)$ symbols of the third kind. For a fixed value of P, $\left\langle \hat{H}\,(J \mid \{v, w\}) \right\rangle$ is only a function of n_v and n_w, and it can be stored in tables. We checked that a proper estimation is obtained using $s_0 = 32$ and taking 500 mixes, with errors below 1%.

References

[1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, **27**, 1948 pp. 379–423 and 623–625.

[2] A. N. Kolmogorov, "Three Approaches to the Quantitative Definition of Information," *Problems of Information Transmission*, **1**(1), 1965 pp. 1–7.

[3] A. Lempel and J. Ziv, "On the Complexity of Finite Sequences," *IEEE Transactions on Information Theory*, **22**(1), 1976 pp. 75–81. doi:10.1109/TIT.1976.1055501.

https://doi.org/10.25088/ComplexSystems.24.2.127

[4] W. Li, "Mutual Information Functions versus Correlation Functions," *Journal of Statistical Physics*, **60**(5–6), 1990 pp. 823–837. doi:10.1007/BF01025996.

[5] M. A. Nowak, N. L. Komarova, and P. Niyogi, "Computational and Evo lutionary Aspects of Language," *Nature*, **417**, 2002 pp. 611–617. doi:10.1038/nature00771.

[6] E. Lieberman, J. B. Michel, J. Jackson T. Tang, and M. A. Nowak, "Quantifying the Evolutionary Dynamics of Language," *Nature*, **449**, 2007 pp. 713–716. doi:10.1038/nature06137.

[7] C. E. Shannon, "Prediction and Entropy of Printed English," *Bell System Technical Journal*, **30**(1), 1951 pp. 50–64. doi:10.1002/j.1538-7305.1951.tb01366.x.

[8] P. Grassberger, "Estimating the Information Content of Symbol Sequences and Efficient Codes," *IEEE Transactions on Information Theory*, **35**(3), 1989 pp. 669–675. doi:10.1109/18.30993.

[9] W. Ebeling and T. Pöschel, "Entropy and Long-Range Correlations in Literary English," *Europhysics Letters*, **26**(4), 1994 pp. 241–246. doi:10.1209/0295-5075/26/4/001.

[10] E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, and E. Moses, "Hierarchical Structures Induce Long-Range Dynamical Correlations in Written Texts," *Proceedings of the National Academy of Sciences of the United States of America*, **103**(21), 2006 pp. 7956–7961. doi:10.1073/pnas.0510673103.

[11] E. G. Altmann, G. Cristadoro, and M. D. Esposti, "On the Origin of Long-Range Correlations in Texts," *Proceedings of the National Academy of Sciences of the United States of America*, **109**(29), 2012 pp. 11582–11587. doi:10.1073/pnas.1117723109.

[12] M. A. Montemurro and D. H. Zanette, "Towards the Quantification of the Semantic Information Encoded in Written Language," *Advances in Complex Systems*, **13**(2), 2010 pp. 135–153. doi:10.1142/S0219525910002530.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Hoboken, NJ: Wiley-Interscience, 2006.

[14] R. Ferrer i Cancho and R. V. Sol, "Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited," *Journal of Quantitative Linguistics*, **8**(3), 2001 pp. 165–173. doi:10.1076/jqul.8.3.165.4101.

[15] E. Ravasz and A.-L. Barabási, "Hierarchical Organization in Complex Networks," *Physical Review E*, **67**, 2003 026112. doi:10.1103/PhysRevE.67.026112.

[16] Project Gutenberg. www.gutenberg.org.

https://doi.org/10.25088/ComplexSystems.24.2.127