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

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

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 $\Delta I_{\{w\}}(s)$).

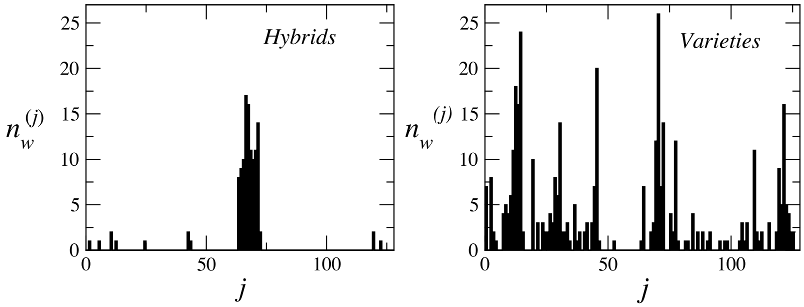


Figure 2. Number of occurrences $n_w^{(j)}$ as a function of the parts j of size $s = 1182$ of the book *On the Origin of Species* for the words *hybrids* and *varieties*.

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

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

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

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

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

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

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

We observe that the probability $p\left(j \mid \{v, w\}\right)$ 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 $H\left(J \mid \{v, w\}\right) \sim \log_2\left(m\right)$. So basically, $\Delta I_{\{v,w\}}\left(s\right)$ 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 $\left\langle \hat{H}\left(J \mid \{v, w\}\right)\right\rangle$ 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

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

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

same way, it can be shown that $\left\langle \hat{H}\,(J \mid \{u, w\})\right\rangle \simeq \left\langle \hat{H}\,(J \mid w)\right\rangle$ 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

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

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

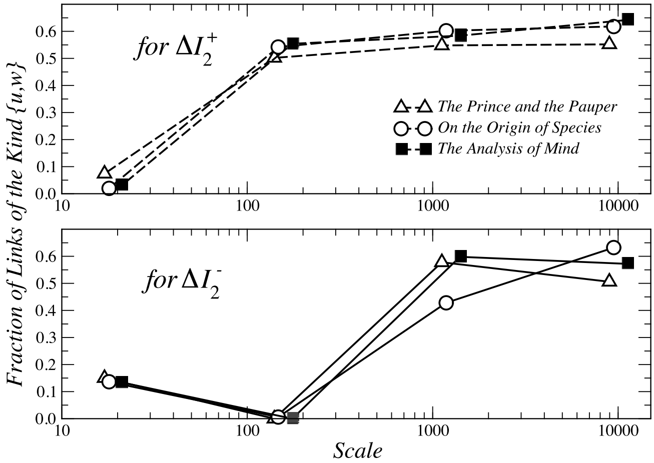


Figure 4. Fraction of links of $\Delta\, I_2\,(s)$ composed of a high-frequency word and a word from $\Delta\, I_1\,(s)$ as a function of the scale s. We consider the first 500 links ranked by their contribution to $\Delta\, I_2^\pm$, the first 100 words from $\Delta\, I_1^\pm$, and the five most frequent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

$p(j \mid \{v, w\}) \propto n_v^{(j)} n_w^{(j)}$ 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.

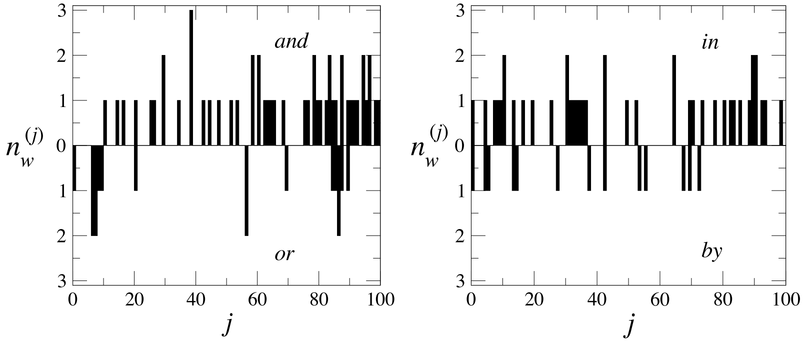


Figure 5. Number of occurrences $n_w^{(j)}$ as a function of j for the first 100 parts of size $s = 18$ of the book *On the Origin of Species* for the pairs $\{$*and*, *or*$\}$ and $\{$*in*, *by*$\}$ (the occurrences of the second word of the pair is plotted as negative for a better view).

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*$\}$.

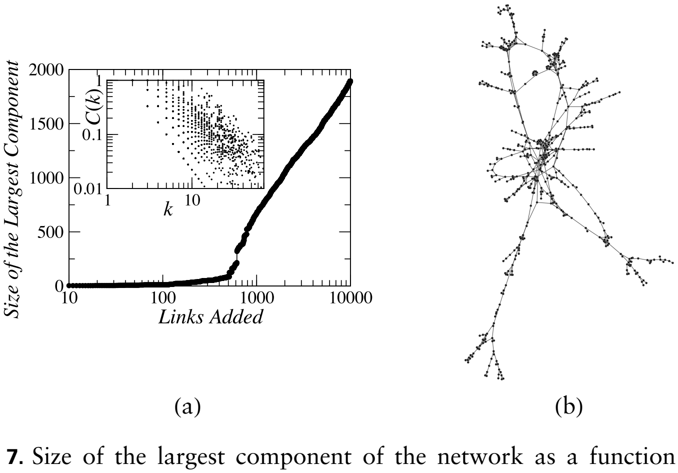
Figure 6 shows $n_w^{(j)}$ as a function of j for the parts of size $s = 147$ of the book *On the Origin of Species* for the words $\{$*pollen*, *flower*$\}$ and for the words $\{$*water*, *fresh*$\}$. In this scale, the text is divided into $P \sim 10^3$ parts, so these words, which have $n_w \lesssim 10^2$, 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

4.1 Network of Words

From the list of links ordered by their contribution to $\Delta I_{\bar{2}}$, 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 $\Delta I_{\bar{2}}$. 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.



(a)

(b)

Figure 7. Size of the largest component of the network as a function of the links added progressively in accordance with their contribution to $\Delta I_{\bar{2}}$ (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*.

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 ($N \sim 6 \times 10^4$).

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 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.

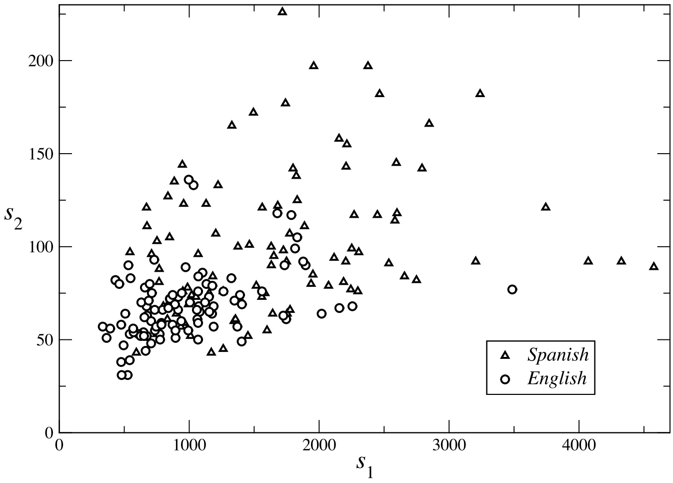


Figure 9. Scale s_1 at which the information ΔI_1 reaches its maximum, and scale s_2 at which ΔI_2 vanishes, for 100 books in Spanish and 100 books in English extracted from Project Gutenberg.

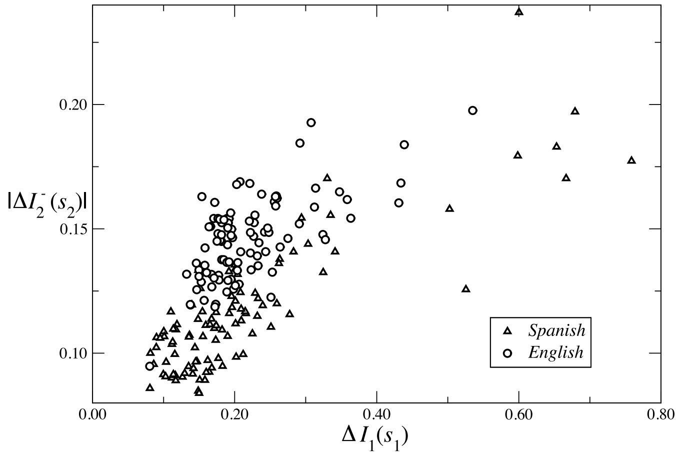


Figure 10. Maximum information per word $\Delta I_1(s_1)$ and the negative component of the information between parts and pairs of words $|\Delta I_2^-(s_2)|$ at the scale where ΔI_2 vanishes, for books in Spanish and English extracted from Project Gutenberg.

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.

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,

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

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 $\langle \hat{p}(j \mid w) \rangle \approx 1 \big/ P$ and the entropy is

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

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

$$\langle \hat{H}\,(J\,|\,w)\rangle \approx \log_2(n_w). \tag{A.6}$$

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

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

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

$$\langle \hat{H}\,(J\,|\,\{v,\,w\})\rangle \approx \langle \hat{H}\,(J\,|\,w)\rangle. \tag{A.8}$$

In this way we just consider $(s_0\,P-n_v-n_w)$ symbols of the third kind. For a fixed value of P, $\langle \hat{H}\,(J\,|\,\{v,\,w\})\rangle$ 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.

[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