The statistics of meaning: Darwin, Gibbon and Moby Dick

How concentrated is the information in Darwin's *Origin of Species*? How heavy a read is Gibbon's *Decline and Fall*? Where does a great white whale fit in? And what on earth is the mysterious Voynich manuscript that has baffled scholars for centuries? **Marcelo Montemurro** and **Damián Zanette** look at words and the information they contain.

It has become a commonplace that computers encode information in the form of sequences of ones and zeros. Step back a few decades, to what went before computers and you realise that books, writing and indeed human language itself encode information in the form of sequences of words.

Any speech or text, from the simplest phrase to the most elaborate volume, is built up from chains of words that follow one another according to a set of rules and constraints that operate at different scales. At the level of individual sentences, word sequences obey the rather strict rules of grammar. In a given message these rules apply independently of the kind of information it conveys. The sentences will obey the same rules whether they make up an instruction manual for a video machine or Tolstoy's *War and Peace*. However, at longer scales linguistic structures arise from the organisation of specific contents. The large-scale structure of, say, a science textbook will be very different from that of a novel or of a diary and all three will differ from popular or heavyweight analyses of history. These large-scale structures will reflect the author's choice of form, style and meaning. Such organisational patterns are primarily determined not by the words themselves, but by the way in which those words are distributed along the message.

Consider, for instance, *On the Origin of Species* by Charles Darwin. In this book, the great naturalist expounds the synthesis of his lifelong observations and thoughts about the realm of living beings that was to set a revolution in our understanding of how life evolved on Earth. In *On the Origin of Species*, the word *islands* appears 143 times. More than 100 of those occurrences come, not unexpectedly, in the two chapters where Darwin discusses the effect of geographical isolation on the appearance of new species. The top panel of Figure 1 shows (coloured line) the distribution of the word *islands* over the text. The higher the curve the larger the number of times the word is used in that particular section of the

All human language shares a coded pattern to express ideas. Can the pattern decipher unknown tongues?

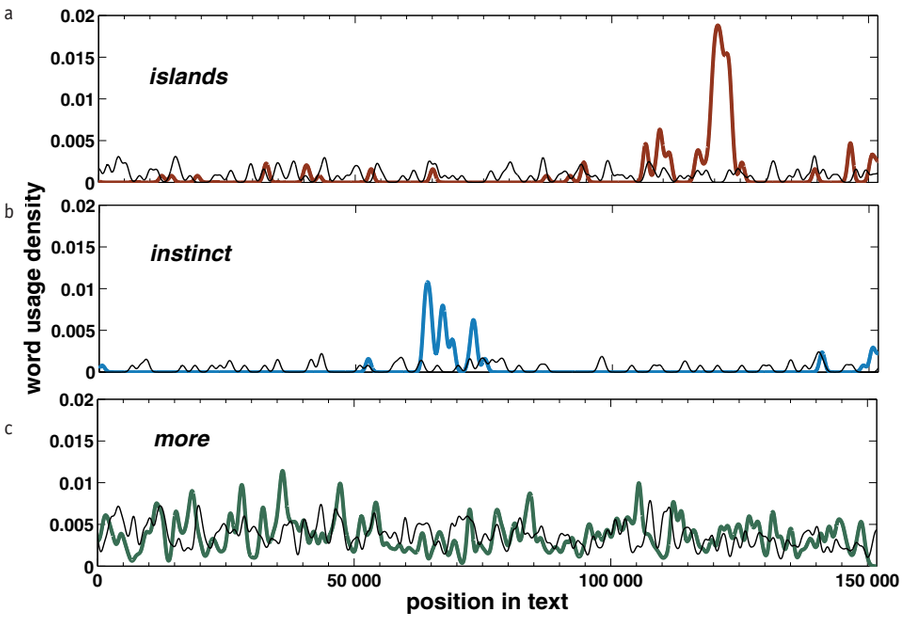


Figure 1. Distribution of three words in *On the Origin of Species*, by Charles Darwin. The coloured curves show the density of word usage for the words *islands*, *instinct* and *more* as a function of the position in the text measured in number of words from the beginning. The higher the density the more frequent is the use of the word around a given position in the text. The black lines correspond to the distribution of the same words after random shuffling of all words in the text

text. A similar pattern is found for the word *instinct*, shown in the middle panel of Figure 1: of its 67 occurrences, more that 50 belong to Chapter 7 whose title is, precisely, "Instinct". However, the word *more*, which appears 576 times, displays a much more homogeneous distribution throughout the text. While *islands* and *instinct* are specific to certain sections of the book, and their meaning is thus expected to bear a close relation with the contents of those sections, the homogeneity of *more* is evidence of its marginal semantic role. In Darwin's work, the words *islands* and *instinct* carry much more meaning, or weight or importance—or information—than the word *more*.

It may seem natural that the way in which words are arranged—those patterns that concentrate the word *islands* into two chapters—is affected by the process of encoding information during the production of a text. However, it is much less obvious that the converse is true: that meaning can be inferred from the pattern of words. Yet this is so. It is possible because of the subtle but systematic relationship between the statistics of word usage and the individual role of words in building up the meaning of the text.

The appropriate mathematical framework to quantify the relationship between statistics of words and the information they convey in language sequences is information theory, which was devised by Claude Shannon in the 1940s. Its original prime application was quantifying electrical transmission of data. We applied information theory to quantify to what extent the specific distribution of words in a text can be used to discriminate between its different sections. Under this rigorous theoretical frame we found a connection between meaning and the statistics of word use.

Note that this is not just the obvious connection between meaning and the actual words used; it is not the words themselves but the *patterns* of word use that give meaning. If you took all the words from *The Origin* then jumbled them up and wrote them out in a random order they would read as nonsense and would carry no meaning at all.

Our quantitative measure of the information borne by the arrangement of words in a text is based on exactly that idea. We compared the actual word distribution in *The Origin* with that of a random shuffling of the same text. The effect of this shuffling is illustrated by the black curves in Figure 1. They represent the distribution of the words *islands*, *instinct* and *more* in the nonsensical random version of *On the Origin of Species*. The crucial difference between the real text and its random counterpart is that in the former the arrangement of words is constrained to encode a linguistic message, whereas the latter lacks any meaning beyond that of individual words. Thus we are quantifying the degree of linguistic order in the text.

Important words tend to clump in paragraphs and chapters that deal with the topic they relate to, but this only provides a crude guide. Using information theory we can go further. We can find the "entropy" of a word and the information content of each word as well. The "entropy" of a word is a measure of how evenly distributed it is. (This is similar to the entropy of a gas in physics—the parallels are almost exact.) We calculate entropies in both the original text and the scrambled version. The difference between the two entropies multiplied by the frequency of the word gives the word's "information value" in the text. Information value, just as in binary computing, is measured in bits.

At the start we explained how dividing *On the Origin of Species* into chapters shows how words like *islands* concentrated themselves and acquired their differing importances. But chapters might not be the best unit to do this: sections of greater or lesser word length might give more meaningful results.

To investigate this, and to calculate the maximum information that the distribution of words can give us, we repeated the analysis many times, each time dividing the text into contiguous parts of equal sizes. The size of each part, measured in number of words, defines a scale, which our information measure compares with the typical scales of word distributions.

In Figure 2 we show the information in bits per word as a function of the scale. We do

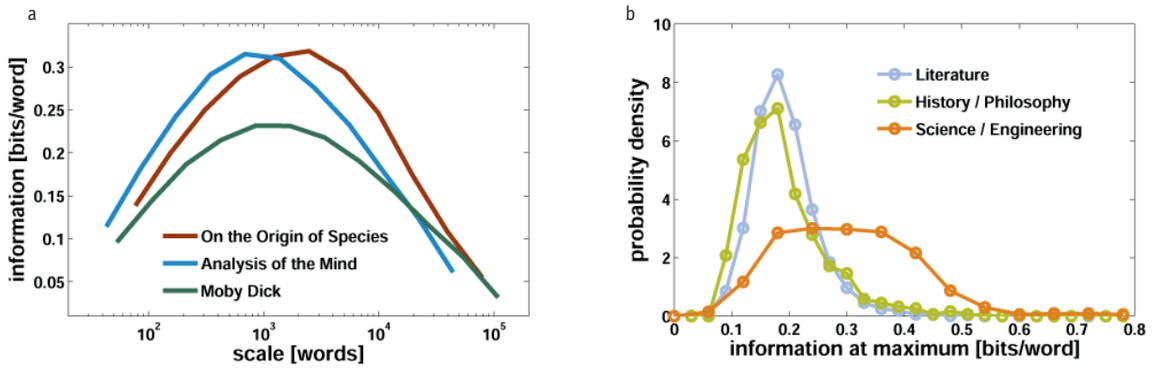
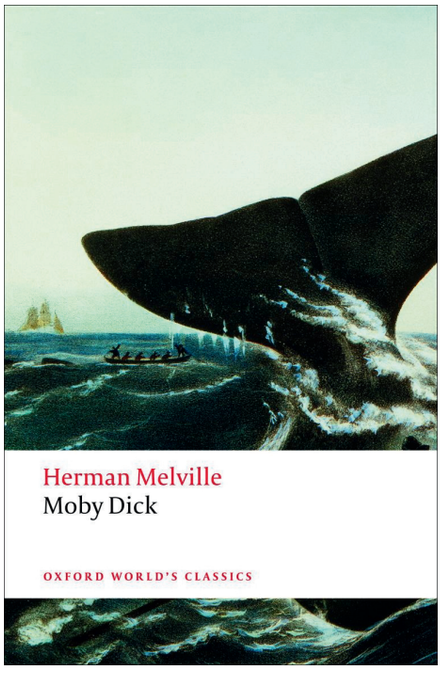


Figure 2. Information encoded in whole texts. (a) Information in bits per word for three texts, as a function of the scale. The texts are *On the Origin of Species* by Charles Darwin, *Analysis of the Mind* by Bertrand Russell and *Moby Dick* by Herman Melville. The scale is the size of each of the parts in which the text is divided to compute the information measure. (b) Normalised histograms showing the distribution of the maximum total information for three subsets of the corpus of English books. The number of books in each division of the corpus was as follows: literature, 3329; history and philosophy, 1374; science and engineering, 555. The distribution of the maximum information within each of the three sub-corpora shows a different profile. This indicates that information theory can quantify the differences in the statistical use of words corresponding to different general topic classes

this for three books: *On the Origin of Species* by Charles Darwin, *Analysis of the Mind* by Bertrand Russell and *Moby Dick* by Herman Melville. The lengths of the texts are, respectively, 155 800, 89 586 and 218 284 words. The most apparent feature in these plots is that the information attains a maximum for a given scale. In these examples, the maximal information occurs at a scale of approximately 2500 words for *On the Origin of Species*, 700 words for *Analysis of the Mind* and 1200 words for *Moby Dick*. This scale is much longer than the range at which grammatical rules apply. It is more related to the text lengths needed to build up meaningful structures beyond the sentence level.



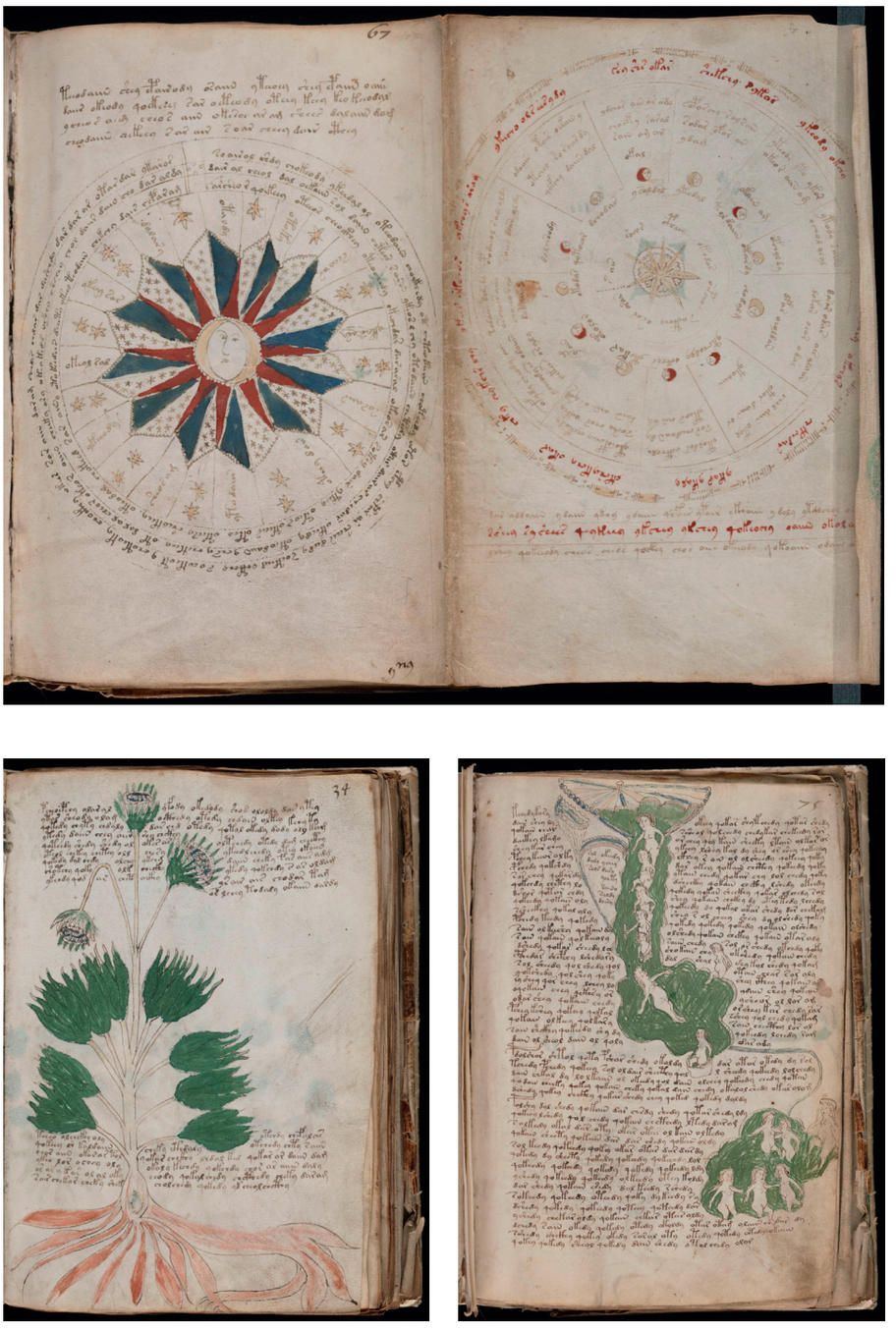
More technically, the presence of an optimal scale means that the distribution of words best discriminates between different sections of the text for a characteristic size of the parts. It is for this particular size that the mutual information of the word distribution and the text partition differs maximally between the real and the random texts.

We have applied this analysis not just to these three books but to thousands of others—and the differences between books—or between *types* of book—are revealing. Our corpus consists of more than 5000 books in English (see Project Gutenberg: `www.gutenberg.org`), including subjects such as English and American Literature, Science, Engineering, Philosophy, History and Religion. Systematically, all the texts presented the same patterns that were seen for the three books analysed in Figure 2, with a maximum of information per word for a given optimal scale. The size of this scale varied between different books, depending on the text lengths required to develop and expose the relevant thematic units in each book. For instance, there were short books with a small optimal scale, such as *Quotations of Lord Chesterfield* and *Quotes and Images from the Novels of Georg Ebers*, which are collections of short quotations with no thematic unity building up along the whole text. Consistently, the optimal scale associated with these books lies in the range between 50 and 70 words. At the opposite extreme, large optimal scales are found in long treatises on subjects with clear thematic unity. The three works with the largest optimal scale, in the range between 4000 and 6000 words, were *History of the Decline and Fall of the Roman Empire, Volume 3* by Edward Gibbon, *A History of Rome, Volume 1* by Abel Greenidge and *Civilization of the Renaissance in Italy* by Jacob Burckhardt. You might care to note that all three are heavyweight works of history, immensely popular in their time (and of course Gibbon, at least, is still read today); and all three are tomes from previous centuries. Burckhardt published in 1860, Greenidge in 1904 and Gibbon's six volumes came out between 1776 and 1788. It would be interesting to find a modern author with a similar scale of vision.

Another interesting aspect is that the maximum value of the information per word for all the books in the collection was narrowly distributed around 0.2 bits/word, despite the large variety of subjects, styles and book lengths. However, when looking at the books within a given range of values of the maximum information per word we found a tendency for literary and history books to have lower values of information per word than many of the books on science and engineering. This can be seen in Figure 2b where we show normalised histograms of the maximum information computed on three subsets of the whole corpus. Whereas the information for the literary, history and philosophy books have a very similar distribution, the values for science and engineering books have a broader distribution that extends over higher values of the information. This shows that information theory can quantify the similarities and difference in language styles in different types of book. In particular, it indicates that the use of language in scientific and engineering books is such that the distribution of words tags the different parts of the text more efficiently.

The Voynich manuscript

The Voynich manuscript is a mysterious, undeciphered illustrated book of some 240 pages, thought to date from the 15th or 16th century. It is hand-written, in an unknown language, and in an unknown alphabet. It is copiously illustrated, with pictures of unidentified plants and what appear to be cosmological and medical diagrams; yet its origin, purpose and meaning have baffled scholars and defied cryptographers for centuries. Scholarly opinion tends to believe that it is not a hoax. It has been suggested that the language is a European one, disguised through several layers of complex secret coding; or a made-up one, created for that volume only; or an Asian language, such as 15th-century Tibetan or Burmese, written straightforwardly but in a specially-invented script. It has been claimed that statistical analysis of word repetitions and word-lengths support the latter view though the word-entropy, of around 10 bits per word, is similar to that of English or Latin.



Pages from the Voynich manuscript. The "biological" section of the manuscript has dense text and illustrations showing nude women bathing. An astronomical section contains circular diagrams, some containing conventional signs of the zodiac but many completely obscure

Finally, we provide evidence that the information being captured with our method is related to the semantic role of words—that is, their role in giving meaning. The information of the whole text is made up of the sum of contributions from every word in the text. Each of these contributions can be interpreted as the information encoded by each individual word. Therefore, we computed the single-word information for all the words in *On the Origin of the Species*, in *Analysis of the Mind* and in *Moby Dick*, at the scale where the total information was maximal for each of the texts (see Figure 2A). Table 1 presents the first 20 words of the three texts ranked by the information of each word. They are not necessarily among the most frequent words in the book, which do not carry any meaning about the content of the book and are just functional words—such as *the*, *and*, *of* or *a*. They carry no information because, in real texts, the most common words have distributions that do not differ much from that of a typical realisation of the randomly shuffled text. Remarkably, however, most of the words with largest information are specifically relevant to the main subjects of the text. Among the top ten words of *On the Origin of Species*, for instance, we find *species*, *varieties*, *hybrids*, *forms*, *islands*, *selection* and *genera*. Those terms would cer-

Table 1. Most informative words for three books

On the Origin of Species	*Analysis of the Mind*	*Moby Dick*
on	image	I
species	memory	whale
varieties	images	you
hybrids	word	Ahab
forms	belief	is
islands	words	ye
of	desire	Queequeg
will	sensations	thou
selection	object	me
genera	you	of
plants	past	he
seeds	knowledge	captain
sterility	box	boat
fertility	content	the
characters	consciousness	Stubb
breeds	appearances	his
groups	movements	Jonah
water	mnemic	was
the	feeling	whales
formations	proposition	my

List of the first 20 most informative words, ordered by decreasing information for the books listed. For each case, all the information values were estimated at the scale that maximised the total information (corresponding to the maximum of the information in Figure 2a).

tainly be singled out by a reader as some of the most representative of the message conveyed by the text. In *Analysis of the Mind* we recognise a very similar pattern. Its top words are essential to the philosophical subject of the book. The comparison with *Moby Dick* is interesting because, it being a novel, its style is very different to that of Darwin's and Russell's treatises. As in other books of its kind, much of the content is built up around its characters, through a network of relationships that change throughout the text; this is evidenced in the list of Table 1 by the prominence of pronouns and proper names in addition to the nouns that set the thematic focus of the book.

We have analysed texts in other languages as well and found similar features. Since the type of information that is being captured is associated with scales far beyond the scope of grammatical rules, a certain degree of universality is expected across languages. And indeed long-range linguistic patterns did appear that seem common to all languages. Probably the underlying mechanisms that humans use to structure complex meaning into strings of words obey general cognitive constraints rather than particularities of each tongue. If that is the case, one can envision exciting applications of these methods to the analysis of old and extinct languages—even to languages still undeciphered. Motivated by that possibility, we have recently started to investigate the relationship between statistics and information in the distribution of words in the Voynich manuscript (`http://en.wikipedia.org/wiki/Voynich_manuscript`; see box). This mysterious medieval manuscript has eluded any attempt to decode it thus far.

Our methods can throw light on many standing issues of this kind. In particular, they can distil from a text a core vocabulary that characterises its overall semantic content. By combining our approach with other advanced statistical methods it may be possible to gain better insight into its linguistic structure.

In a broader framework, the type of analysis we have reviewed could be extended to study other natural information carriers such as neural signals, the genetic code and patterns of animal communication and behaviour. All these examples of communication systems can be mapped into strings of symbols and thus our approach is directly applicable. Overall, our results suggest that complex aspects of the information encoded in symbolic sequences are susceptible to quantitative characterisation and analysis using the rigorous principles of information theory[1].

Reference

1. Montemurro M. A. and Zanette D. H. (2009) Towards the quantification of the semantic information encoded in written language. Preprint available at http://arxiv.org/abs/0907.1558

Marcelo Montemurro is MRC Research Fellow in Bioinformatics Neuroinformatics at the Faculty of Life Sciences at the University of Manchester. Damián Zanette researches non-equilibrium statistical physics at the Consejo Nacional de Investigaciones Científicas y Técnicas, Centro Atómico Bariloche and Instituto Balseiro, Río Negro, Argentina.