Statistical Patterns in Written Language

Damián H. Zanette

Statistical Patterns in Written Language

Damián H. Zanette
Centro Atómico Bariloche, Argentina
zanette@cab.cnea.gov.ar, http://fisica.cab.cnea.gov.ar/estadistica/zanette/

Version 1.1 – September 2012

Contents

Preface

Quantitative linguistics has been allowed, in the last few decades, within the admittedly blurry boundaries of the field of complex systems. A growing host of applied mathematicians and statistical physicists devote their efforts to disclose regularities, correlations, patterns, and structural properties of language streams, using techniques borrowed from statistics and information theory. Overall, results can still be categorized as modest, but the prospects are promising: medium- and long-range features in the organization of human language —which are beyond the scope of traditional linguistics— have already emerged from this kind of analysis and continue to be reported, contributing a new perspective to our understanding of this most complex communication system.

This short book is intended to review some of these recent contributions. To articulate them within a coherent context, however, some space is devoted to summarize some classical work on the same subjects, dating back to the middle decades of the twentieth century. Specifically, the topics discussed here are directly related to —and frequently derived from— the seminal quantitative studies of language by George Zipf, Claude Shannon, and Herbert Simon.

After an introductory chapter on probability, statistics, and information theory —included for the benefit of the reader who may not be closely familiar with the mathematical background relevant to the remaining of the book— the first topic refers to the frequency of word usage in written language. The universal statistical regularity known as Zipf's law and its derivations are reviewed in the light of classical and recent work, discussing their connection with the emergence of context, and their occurrence in several phenomena outside linguistic systems. The second main subject regards long-range correlations in language streams, as detected in actual character and word sequences. Burstiness in the distribution of highly topical words, which may provide a way to automatically identify keywords in any given text, is discussed in connection with those correlations. Finally, the intriguing problem of assigning a quantitative measure to the information contained in language —first addressed by Claude Shannon when founding information theory— is reviewed together with a few derived issues, such as the contribution of word ordering to informational contents, and the definition

of semantic length scales in language samples.

Necessarily, the selection of subjects is biased and, not unexpectedly, it turns out to be highly correlated with my own work in the field —most of it, done in collaboration with my life-long friend and colleague Dr. Marcelo Montemurro, from the University of Manchester. I take this opportunity to acknowledge Marcelo's preeminent role in the achievement of our joint contributions.

Damián H. Zanette
San Carlos de Bariloche, September 2012

Chapter 1

Basic Concepts of Statistics and Information Theory

Human language evolved —and continues to change— under the pressure for improving efficiency in its primary function, namely, the exchange of meaningful information. Multiple factors enter this evolution and affect each other in intricate ways, from the anatomic transformations of the vocal tract and the brain's centers of speech and hearing, to the consensual emergence and adaptation of grammar rules and semantic order, which organize the components of language into intelligible messages. At present, this complex optimization process is at an advanced stage, so that our everyday use of language leaves relatively little room for the introduction of fortuitous elements or random constructions, which would imperil the success of communication. Only certain tongues, within specific literary genres such as poetry, admit a moderate level of arbitrariness on the basis of aesthetic criteria.

From the viewpoint of traditional linguistics, thus, it may come as a surprise that mathematical approaches such as statistics and information theory —which are purposedly designed to deal with systems driven, in large measure, by random mechanisms— have proven fruitful in the quantitative analysis of the structure of language. The reason for this success, however, is not difficult to grasp from a broader perspective: the complexity of linguistic patterns, which is inherent to communication between humans and deploys over multiple scales and at many levels along language streams, admits a statistical description much in the way as other complex natural systems, even those governed by deterministic laws. This book reviews a series of contributions within such approach, which —since the first half of the twentieth century, but specially in the last few decades— disclosed statistical regularities in the frequency of word usage, in the unfolding of semantic contents over long texts, and in the emergence of topicality and context.

The purpose of the present chapter is to offer a swift introduction to the mathematical concepts relevant to the topics discussed in Chaps. 2 to 4, for the benefit of the reader who may be not too familiar with the basic ideas and methods of probability, statistics, and information theory (but from whom we require a minimum of acquaintance with elementary mathematical notation). The presentation is not mathematically formal, and strongly focuses on the concepts useful to follow the remaining of the book. Examples with applications to the analysis of language samples are given, as preliminary clues to the use of such methods in the context of linguistics. For more comprehensive presentations of statistics and information theory, the handbooks by Gardiner (2004) and Cover and Thomas (2006) are respectively recommended. The reader who is well acquainted with statistical techniques, and who understands why they can be usefully applied in the field of quantitative linguistics, can safely skip this chapter and proceed directly to Chap. 2.

1.1 Probability: Definitions and Interpretation

The introduction of the notion of probability presupposes the existence of a set of events, to which probabilities are to be assigned (Gardiner, 2004). These events occur as the outcome of a given process under specified conditions, but the mechanisms which govern the process are not fully accessible to the observer. Therefore, some degree of unpredictability —and, hence, the emergence of random elements in the description of the process— is assumed. A typical example of this scenario is an experiment with a prescribed protocol, whose outcome belongs to the set of all possible results. For instance, the tossing of a coin is associated with two events, corresponding to the two possible results: head or tail. For the sake of concreteness, we refer in the following discussion to this kind of probabilistic experiments and their results.

The probability assigned to each result is a quantitative measure of the likelihood that the result effectively occurs as the outcome of the specified experiment. For our present purposes, it is convenient to assume that results are countable, so that each of them can be denoted by r_i, with $i = 1, 2, \ldots, K$, where K is the total number of possible results. Let us also assume that results are defined in such a way that they do not overlap, namely, that any two of them cannot occur simultaneously —they are mutually exclusive. If $p(r_i)$ is the probability of result r_i, the following axioms are established:
(i) Non-negativity: $p(r_i) \geq 0$ for all $i = 1, 2, \ldots, K$.
(ii) Normalization: $\sum_{i=1}^{K} p(r_i) = 1$.
(iii) Addition rule: the probability that the outcome of the experiment is *either* result r_i *or* result r_j is $p(r_i) + p(r_j)$.

The ensemble of probabilities $p(r_1), p(r_2), \ldots, p(r_K)$ defines the *probability distribution* over the set of results.

In the so-called *frequentist interpretation* of probability theory —which, within the context of the following chapters, provides the appropriate conceptual framework for applications— the probability of a result equals the relative frequency with which that result occurs in an infinitely long series of equivalent realizations of the experiment in question. If, in a series of R realizations, the outcome coincides with result r_i a number R_i of times, we have

$$p(r_i) = \lim_{R\to\infty} \frac{R_i}{R}. \tag{1.1}$$

Assuming that, in the limit, the values of R_i are well defined for all i, the probabilities given by Eq. (1.1) satisfy the above axioms.

The frequentist definition of probabilities has the obvious drawback that, to determine $p(r_i)$ for any result, it is formally necessary to perform infinitely many repetitions of the same experiment. Empirically, therefore, probabilities are in principle accessible up to a certain precision only, due to the natural limitation in the number of repetitions achievable by the experimenter. Under certain conditions, however, probabilities can be evaluated from preexisting information about the nature of processes being considered. To illustrate this, let us consider the following experiment, which is relevant to the statistical approach to written texts addressed in the following chapters. Take a copy of Charles Dickens's *David Copperfield* and open it at random. With your eyes closed, put your finger anywhere on the open page. Now, open your eyes. What is the probability that you are pointing to the word *the*? Instead of repeating the experiment hundreds of thousands of times —Dickens's novel is 363128 words in length, and its lexicon comprises 14078 different words— you may go to Chap. 2 of this book and learn that *the*, which is the most frequent word in *David Copperfield*, occurs there a total of 13763 times. If the book is opened genuinely at random and the finger pointed to a really arbitrary place in the text, and if all realizations of the experiment are absolutely independent of each other so that any mutual influence can be discarded, it is intuitively clear that the probability of pointing to *the* coincides with the ratio between its number of occurrences and the total text length: $p(the) = 13763/363128 \approx 0.038$. This means that, from each thousand repetitions of the experiment, *the* will be the outcome some 38 times.

Obviously, however, a problem of circular definition is enclosed in these arguments. Specifically, the notions of randomness, arbitrariness, and independence involved in the experimental protocol hide the presupposition that all places in the text of *David Copperfield* are equally likely to be chosen — namely, that any of them will be pointed to with the same probability. This aprioristic knowledge of the probabilities of certain elementary events pervades the theory, and persists even in its most formal settings. The intuitive

knowledge of *a priori probabilities* is ubiquitous in any practical application of the formalism or, as C. W. Gardiner (2004) put it, *"there is no way of making probability theory correspond to reality without requiring a certain degree of intuition."*

Of importance in many applications of probability theory are the concepts of *joint* and *conditional* probabilities. These notions involve the coexistence of two (or more) experiments, not necessarily independent of each other. Their interdependence may, for instance, be originated in some common underlying mechanism or in a causal relation to one another. Suppose to have two such experiments, whose possible outcomes are respectively given by the result sets $r_1, r_2, \ldots, r_K$ and $r'_1, r'_2, \ldots, r'_{K'}$. The joint probability of results r_i and r'_j, which we denote by $p(r_i, r'_j)$, is the probability that in a joint realization of the two experiments their respective outcomes are those two specific results. In terms of the joint probabilities, the probabilities of individual results in either experiment —called, in this context, *marginal* probabilities— are given by

$$p(r_i) = \sum_{j=1}^{K'} p(r_i, r'_j), \qquad p(r'_j) = \sum_{i=1}^{K} p(r_i, r'_j). \tag{1.2}$$

By definition, the two experiments are independent if, for every pair of results, their joint probability can be factored out as $p(r_i, r'_j) = p(r_i)p(r'_j)$.

The conditional probability of result r_i given result r'_j, denoted by $p(r_i|r'_j)$, is the probability of obtaining result r_i from the first experiment under the condition that the outcome of the second experiment is r'_j. A similar definition holds for $p(r'_j|r_i)$. Conditional and joint probabilities are linked together by the relations

$$p(r_i, r'_j) = p(r_i|r'_j)p(r'_j) = p(r'_j|r_i)p(r_i). \tag{1.3}$$

The second identity in this equation is known as Bayes's theorem. This theorem is at the heart of an alternative interpretation of probability theory —the so-called *Bayesian interpretation*— where the probability of an event is conceived as the result of a progressive construction of knowledge on its likelihood, out of the aprioristic estimation of the probability of more elementary events, instead of relying on the empirical observation of its frequency.

As an illustration of the relation between joint and conditional probabilities consider the above experiment of choosing a word from *David Copperfield* at random. As the second experiment, suppose that you inspect the word which immediately follows, along the text, the word chosen in the first experiment, and determine whether it is a noun or not. If the outcomes of a joint realization of the two experiments are, respectively, the word *the* and a noun, Eq. (1.3) establishes that

$$p(the, noun) = p(the|noun)p(noun) = p(noun|the)p(the). \tag{1.4}$$

We have already found that $p(\textit{the}) \approx 0.038$. If we assume that, because of its grammatical function as an article, the word *the* is unavoidably followed by a noun, the conditional probability of finding a noun given that *the* has been chosen first must express this certainty, namely, $p(\textit{noun}|\textit{the}) = 1$. From the second identity in Eq. (1.4), it follows that the joint probability in the left-hand side of the same equation is $p(\textit{the}, \textit{noun}) \approx 0.038$. If, moreover, we are informed that about one sixth of all words in *David Copperfield* are nouns, i.e. that $p(\textit{noun}) \approx 0.17$, the first identity in Eq. (1.4) allows us to conclude that $p(\textit{the}|\textit{noun}) \approx 0.038/0.17 \approx 0.23$. In other words, every 100 joint realizations of the two experiments where the outcome of the second experiment is a noun, some 23 realizations have *the* as the outcome of the first.

1.2 Random Variables and Stochastic Processes

It is often the case that the outcome of a probabilistic experiment, as those considered in the preceding section, is associated with a measurable quantity x. In this situation, it is useful and customary to characterize the outcome of the experiment by the set of all the different values, $x_1, x_2, \ldots, x_N$, that the measurement of the quantity x can return.[1] Using the frequentist definition, Eq. (1.1), the relative frequency of occurrence of x_i in the measurement of x gives the probability $p(x_i)$. The set $p(x_1), p(x_2), \ldots, p(x_N)$ constitutes the probability distribution over the quantity x, and x becomes thus defined as a *random variable*.

As an example of a random variable, recall from the preceding section the experiment of choosing a word at random from the text of *David Copperfield*, and define x as the number of letters in the chosen word. The smallest value of x as a random variable in this experiment is 1, as in the words *a* and *I*, with a probability $p(1) \approx 0.07$. The largest is 29, with $p(29) \approx 2.75 \times 10^{-6}$, standing for the only occurrence of the "word" *retheguidingstarofmyexistence* by the middle of the novel's twenty-fourth chapter. The maximal probability in the distribution is obtained for three-letter words, with $p(3) \approx 0.22$. No doubt, the very frequent words *the* and *and* (see Table 2.1 in Chap. 2) contribute substantially to this probability. Figure 2.7 shows the probability distribution for word lengths in *David Copperfield*, up to $x = 16$, along with those for *Don Quijote* and *Æneid* (in Spanish and Latin, respectively).

While a complete statistical characterization of the measurement of a random variable is achieved by fully specifying its probability distribution, useful information is already contained in a smaller set of conveniently defined quantities, which measure different properties of the distribution. The

[1] Note that there is no need that the different values of x and the different results of the experiment, $r_1, r_2, ..., r_K$, are linked by a bijective relation. A single value of x may correspond to two or more results.

mean value or *average* of the random variable x is defined as

$$\langle x \rangle = \sum_{i=1}^{N} p(x_i) x_i. \tag{1.5}$$

In formal probability theory, $\langle x \rangle$ is often called the *expectation value* of x, and denoted by $E(x)$. Another common notation for the mean value of x is $\bar{x}$. The mean value of a random variable quantifies the typical measurement result. From the frequentist definition of probabilities, Eq. (1.1), it can be shown that $\langle x \rangle$ coincides with the arithmetic average of the results in a sufficiently long series of measurements of x. In the above example of word lengths in *David Copperfield*, it turns out that $\langle x \rangle \approx 4.06$, while for *Don Quijote* and *Æneid* we get $\langle x \rangle \approx 4.31$ and 5.76, respectively. On the average, thus, Latin words seem to be considerably longer than their English and Spanish counterparts.

The notion of mean value is straightforwardly extended to any function $f(x)$ of the random variable, as

$$\langle f(x) \rangle = \sum_{i=1}^{N} p(x_i) f(x_i). \tag{1.6}$$

This quantity represents a typical value of the function $f(x)$ in the measurement of x. In fact, $\langle f(x) \rangle$ coincides with the arithmetic average of $f(x)$ evaluated over the results of a sufficiently long series of measurements of the random variable.

The mean value of the squared difference between the variable and its average,

$$\sigma_x^2 = \langle (x - \langle x \rangle)^2 \rangle = \sum_{i=1}^{N} p(x_i)(x_i - \langle x \rangle)^2 = \langle x^2 \rangle - \langle x \rangle^2, \tag{1.7}$$

defines the *variance* of x. Its square root, σ_x, quantifies the average distance of the random variable from its mean value, irrespectively of whether x is above or below $\langle x \rangle$, and is called *mean square dispersion* or *standard deviation*. It measures the width of the distribution $p(x_i)$. From the distributions depicted in Fig. 2.7, it turns out that the standard deviations of the word lengths in *David Copperfield*, *Don Quijote*, and *Æneid* are, respectively, $\sigma_x \approx 2.28$, 2.52, and 2.21. It is interesting that the Latin text —which, as we have seen above, has the longest words on the average— exhibits the lowest dispersion in word length.

The last identity in Eq. (1.7) shows that the variance can be written in terms of the mean value of x and of its square x^2. Higher-order moments of the probability distribution, $\langle x^k \rangle$ for $k = 3, 4, \ldots$, can be combined to characterize further features of the distribution such as, for instance, its

symmetry or self-similarity properties (see Sec. 3.1). Their usefulness in a specific problem depends on the degree of detail with which the probability distribution needs to be described.

The statistical measures introduced above to quantify the probability distribution of a random variable x admit a straightforward extension to the case in which, in contrast with the assumption made so far, x varies continuously over a certain numeric interval (a, b). In this situation, however, it is necessary to generalize the concept of probability to continuous variables. The probability distribution $p(x)$ of the random variable $x \in (a, b)$ is defined as a function such that the product $p(x)dx$ is the (infinitesimally small) probability that the variable adopts a value within an interval of length dx around x. The normalization of probability (see Sec. 1.1) requires that

$$1 = \int_a^b p(x)dx. \tag{1.8}$$

The mean value of a function $f(x)$ defined over (a, b) is

$$\langle f(x) \rangle = \int_a^b p(x)f(x)dx. \tag{1.9}$$

This expression generalizes Eq. (1.6), and makes it possible to extend the definition of mean value and mean square dispersion of the random variable, Eqs. (1.5) and (1.7), to continuous domains.

Many applications of probability theory involve considering ordered sequences of random variables resulting, for instance, from successive realizations of an experiment along time. If the variable is recorded at discrete steps, coinciding with times $t_1, t_2, t_3, \ldots$, the sequence can be denoted as $x(t_1), x(t_2), x(t_3), \ldots$, where $x(t_i)$ is the value of x measured at step i. Such a time-dependent random variable is called a *stochastic process*. A full statistical characterization of a stochastic process would be in principle achievable by specifying a joint probability for the occurrence of each possible value of x at each time in the sequence of measurements. This quantity, however, is seldom available: either empirical data or a model of the mechanisms that drive the process provide, at most, partial information on how successive realizations of the random variable influence each other.

As an example of a stochastic process consider the following *random walk* (see also Secs. 3.1 and 3.2): a particle moves along a line, jumping at each time step —each second, say— a fixed distance d forwards or backwards, with the same probability in either direction: $p(\textit{forwards}) = p(\textit{backwards}) = 0.5$. The direction is chosen independently at each step. The random variable Δ which characterizes the displacement at each jump, whose possible values are $+d$ and $-d$, defines itself a stochastic process $\Delta(t)$. The current position of the random walker, $x(t)$, which results from the accumulation of

all the displacements from its initial position $x(0)$,

$$x(t) = x(0) + \sum_{u=1}^{t} \Delta(u), \tag{1.10}$$

is a stochastic process as well. Its probability distribution, however, needs to be defined in terms of the dynamical rules which govern the process. Specifically, the probability that $x(t)$ adopts a certain value —i.e., the probability that the random walker is at a certain position at time t— depends on what was the walker's position before the last jump. It turns out that, in this random walk, the average position of the walker coincides at all times with its initial position $x(0)$. On the other hand, the mean square dispersion of $x(t)$ —which quantifies the average displacement in either direction from the initial condition— grows proportionally to the square root of time: $\sigma_x(t) = \sqrt{2Dt}$. The coefficient D is the random walker's *diffusivity*, and depends on the jump length d and on the time between jumps.

The time-like dimension of stochastic processes makes it possible to introduce a new set of statistical measures, defined on the basis of temporal averages. Hence, the *time average* of a stochastic process is given by

$$\langle x \rangle_t = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} x(t_i), \tag{1.11}$$

where it is assumed that t_N tends to infinity with N. The same kind of arithmetic average of the successive values of any given function of x, $f[x(t_i)]$, defines the time average $\langle f(x) \rangle_t$. An important statistical characterization of a stochastic process is given by the two-time correlation function

$$c(t_i, t_j) = \langle [x(t_i) - \langle x \rangle_t][x(t_j) - \langle x \rangle_t] \rangle, \tag{1.12}$$

which measures how interdependent are the values of the stochastic process at different times, on the average over different realizations of the process. For $t_i = t_j$, $c(t_i, t_j)$ is the variance of $x(t_i)$ over those realizations. Generally, positive values of $c(t_i, t_j)$ indicate the tendency of the random variable to remain either above or below $\langle x \rangle_t$ as time elapses from t_i to t_j, while negative correlations are obtained when the variable is usually found at different sides of $\langle x \rangle_t$ at those two times. A stochastic process is *stationary* when $c(t_i, t_j)$ depends on t_i and t_j through the time interval $t_j - t_i$ only.

Statistical measures as defined above —which are given by, or coincide with, arithmetic averages either over different realizations of a specified process or over time— establish an intuitive connection between the description of random processes, on one side, and of other observable quantities that may have no inherent randomness, but which are amenable to the same kind of treatment. This connection is at the basis of the application of statistical techniques to the study of language structures which, as discussed

in the introduction to this chapter, are hardly ascribable to any random phenomenon. Recall, for instance, the distribution of the length of words in a literary text. Word lengths depend on such definite factors as etymological history, internal morphology, and scripture. This definiteness does not preclude, however, to calculate the mean value or the variance of lengths using exactly the same expressions as if they where random variables, as we have done above, with the only proviso of identifying probabilities with frequencies and *vice versa*. Even more, the qualitative interpretation of these statistical measures, as typical values of the averaged quantities over the sets under study, is the same as for random variables.

Along the same line of comparison, the sequential nature of human communication —as a succession of spoken utterances and gestures, or of written symbols such as letters, words, or ideograms— suggests a statistical description similar to that of the ordered array of random values coming out from a stochastic process. In written language, the time dimension is replaced by an index to the place of each symbol along the text but, otherwise, the correspondence is transparent. It is from this perspective that linguistic structures are dealt with by means of statistical techniques, and it is on this viewpoint that the presentation of the following three chapters is grounded.

1.3 Entropy, (Dis)order, (Un)certainty, and Information

Introduced by Rudolf Clausius in the 1850s to quantify the heat dissipated by a physical system during a thermodynamic transformation, the notion of *entropy* was later extended by Ludwig Boltzmann and others to encompass out-of-equilibrium processes, as a key concept of statistical mechanics (Huang, 1987). It was however Boltzmann's student Paul Ehrenfest and his wife Tatyana who clarified, by means of a series of cleverly designed models, how entropy could be used as a direct measure of the disorder of a system or, equivalently, as an inverse measure of its degree of organization (Ehrenfest and Ehrenfest, 1911).

Boltzmann defined the entropy of a physical system in terms of the number of states Ω accessible to the system under certain constraints, such as conservation laws, as

$$S = -k \ln \Omega. \tag{1.13}$$

The coefficient k fixes the units of S and is called Boltzmann constant. The Ehrenfests realized that the same definition provided a useful quantification of the (inverse) degree of order of any system —not only in the realm of physical phenomena— that can access a given set of configurations. Consider, for instance, their "urn model," where M balls are distributed among two containers, with M_1 balls in the first container and $M_2 = M - M_1$ balls in the second. If the balls are distinguishable from each other, there are

$\Omega = M!/M_1!M_2!$ ways to perform the distribution. The entropy of the set of balls so distributed is

$$S = -k \ln \frac{M!}{M_1!M_2!} \approx -kM \left(\frac{M_1}{M} \ln \frac{M_1}{M} + \frac{M_2}{M} \ln \frac{M_2}{M} \right), \tag{1.14}$$

where the last approximation is obtained from Stirling's formula when M, M_1 and M_2 are large numbers (Abramowitz and Stegun, 1972). It can be easily seen that the entropy is minimal, $S = 0$, when either $M_1 = 0$ or $M_2 = 0$, i.e. when all the balls are in the same container and the other container is empty —a situation identified with the maximal degree of order in the system. On the other hand, S is maximal for $M_1/M = M_2/M = 0.5$, when the balls are equally distributed among the two containers.

The generalization of the urn model to the case of K containers is straightforward. The entropy of a distribution of M balls divided into K groups of $M_1, M_2, \ldots, M_K$ balls, is

$$S = -k \ln \frac{M!}{M_1!M_2! \cdots M_K!} \approx -kM \sum_{i=1}^{K} f_i \ln f_i, \tag{1.15}$$

with $f_i = M_i/M$ being the fraction of balls in urn i. These fractions satisfy $\sum_{i=1}^{K} f_i = 1$.

In his foundational work on information theory, Claude Shannon (1948a,b) demonstrated that, given a probability distribution $p(r_i)$ $(i = 1, 2, \ldots, K)$ over a set of K events $r_1, r_2, \ldots, r_K$, the quantity

$$H = -\sum_{i=1}^{K} p(r_i) \log_2 p(r_i) \tag{1.16}$$

satisfies a series of desirable mathematical properties for the definition of the *uncertainty* associated with the distribution: it is a continuous function of $p(r_i)$; when all the probabilities are equal, i.e. when $p(r_i) = K^{-1}$, H is an increasing function of K; and when each event r_i can be thought of as the joint occurrence of two independent sub-events r_i^{I} and r_i^{II}, such that $p(r_i)$ is the product of the probabilities of the two sub-events, the function H splits into two contributions corresponding to each set of sub-events, namely, $H = H_{\mathrm{I}} + H_{\mathrm{II}}$. The quantity H is called the entropy of the probability distribution $p(r_i)$. The choice of the base-2 logarithm $\log_2$ defines the units of H, which is thus measured in *bits* (an acronym for "binary units").

As an illustration, we can immediately evaluate from Eq. (1.16) the entropies of the probability distributions of word lengths for *David Copperfield*, *Don Quijote*, and *Æneid*, already considered in the previous section (see also Fig. 2.7). This requires associating each event with the occurrence of a given word length in the process of choosing words at random from each book. The respective results are $H \approx 3.02$, 3.15, and 3.14. As we have seen above,

word lengths in *David Copperfield* and *Don Quijote* have similar mean values. The entropy for the latter is larger because the corresponding standard deviation —and, hence, the dispersion of word lengths— is larger as well. In turn, *David Copperfield* and *Æneid* have similar standard deviations. In this case, the former has a lower entropy because the average word length is also lower and, therefore, the probability distribution is more concentrated toward small lengths.

Shannon exploited the definition of the entropy of a probability distribution to introduce a quantification of the concept of *information*. Information is conceived as the certainty gained (or lost) during a process where our knowledge about a given set of possible events changes, i.e. where the probability distribution varies. As an example, suppose that —under the present weather conditions— we can estimate that it is equally likely that tomorrow it will rain or not, i.e. that $p(rain) = p(no\ rain) = 0.5$. The entropy of this probability distribution is $H_0 = 1$ bit. If we are later informed that the forecast predicts a 70% rain probability, the distribution changes to $p(rain) = 0.7$ and $p(no\ rain) = 0.3$, with $H \approx 0.88$ bits. The entropy has decreased thanks to the forecast, as did our uncertainty about tomorrow's weather. The difference between the entropies,

$$I = H_0 - H, \tag{1.17}$$

is the information gained between the two stages. In our example, $I \approx 0.12$ bits. Note that, in terms of this definition, a bit can be defined as the amount of information gained when the outcome of an experiment with two equally likely events ($H_0 = 1$ bit) —for instance, head or tail in a fair coin tossing— becomes known with certainty. In fact, once the result is known, the entropy equals zero: $H = 0$ bits.

The entropy of a probability distribution, given by Eq. (1.16), inspires the definition of a series of derived entropy- and information-like measures (Cover and Thomas, 2006), useful for various purposes. Of particular relevance for the remaining of this book are the *relative entropy* and the *mutual information*. The relative entropy, or *Kullback–Leibler distance*, is an entropy-like quantification of the distance between two different probability distributions, $p(r_i)$ and $p'(r_i)$ ($i = 1, 2, \ldots, K$), defined over the same set of events. It is given by

$$D(p, p') = \sum_{i=1}^{K} p(r_i) \log_2 \frac{p(r_i)}{p'(r_i)}. \tag{1.18}$$

This quantity measures the difference between the two distributions $p(r_i)$ and $p'(r_i)$: it is always non-negative, and equals zero if and only if $p(r_i) = p'(r_i)$ for all i. On the other hand, unlike standard definitions of distance, it is not symmetric —i.e. $D(p, p') \neq D(p', p)$— and does not satisfy the triangle inequality (Abramowitz and Stegun, 1972).

The mutual information is defined for two sets of events, $r_1, r_2, \ldots, r_K$ and $r'_1, r'_2, \ldots, r'_{K'}$, with joint probabilities $p(r_i, r'_j)$ and marginal probabilities $p(r_i)$ and $p(r'_j)$ given by Eq. (1.2). The mutual information is given by

$$I(r, r') = \sum_{i=1}^{K} \sum_{j=1}^{K'} p(r_i, r'_j) \log_2 \frac{p(r_i, r'_j)}{p(r_i)p(r'_j)}. \tag{1.19}$$

This quantity measures the amount of information that the occurrence of an event from the first set contains about the outcomes from the second, or *vice versa* —in fact, $I(r, r') = I(r', r)$. It provides a kind of information-like correlation between the two sets of events, quantifying the increase of certainty about the events in one of the sets due to knowledge of those of the other.

As it transpires from Shannon's seminal papers, the possibility of measuring the amount of information enclosed in language —with the prospective of comparing different tongues, authors, and styles— was a major stirring force in the development of the mathematical formalization of communication processes. This theoretical framework situated human languages in the broader class of the systems that encode information, together with neural signals, the genetic code, and programming languages, among others. All of them share the multiplicity of organizational levels that emerge from the evolution of their complex function. Getting a fully satisfactory definition for the entropy of language, able to characterize such structural complexity, is still an open problem. As we discuss at various points in the next chapters, it presently attracts much attention in the field of quantitative linguistics.

Chapter 2

Word Frequencies: Zipf's Law and the Emergence of Context

The frequency with which different words are used in writing or in speech —irrespectively of their specific ordering— is arguably the most elementary statistical property of human language. Historically, it has been the first to be quantitatively characterized. In the 1930s, the philologist George Zipf carried out an extensive empirical study of the connection between number of words, usage rates, and word ranking in different kinds of texts and various tongues, and established the mathematical relations that we know today as Zipf's law. Two decades later, the sociologist Herbert Simon proposed a model, based on a multiplicative stochastic process, for the recurrent use of words. Simon's model is able to quantitatively reproduce many of the relations empirically disclosed by Zipf. At the same time, the increasing reinforcement in the usage of certain words at the expense of others, which is the basic mechanism involved in the model, qualitatively explains how context emerges as a text is produced through the author's successive choices, defining style, subject, tense, person, among the various elements that sustain the intelligibility of the message.

In this chapter, we first review Zipf's observations, as well as his and others' attempts to explain the statistics of word usage from the balance between the efforts invested by speaker and hearer in the communication process. Next, we present Simon's model, discussing its interpretation as a toy picture for the emergence of context during text generation, and extending this discussion from language to music, where the concept of context is also a meaningful one. We then consider some extensions of Simon's model, which provide better fittings of linguistic empirical data and make it possible to apply the same kind of description to the dynamics of other systems controlled by multiplicative stochastic processes, such as the demography

of human populations. Finally, we analyze other random-text models for Zipf's law, which motivate revisiting the relevance of Zipf's findings for our understanding of human language.

2.1 Word Frequencies from the Principle of Least Effort

George Kingsley Zipf's *Human Behaviour and the Principle of Least Effort* (Zipf, 1949) is a lively, eclectic discussion on human psychology and sociology, from the viewpoint of a linguist and philologist who spared no effort to demonstrate to the reader that a person will always *"strive to solve his problems in such a way as to minimize the total work that he must expend in solving both his immediate problems and his probable future problems."* Supplying a wealth of quantitative empirical evidence, the book starts by considering the costs of human communication, language acquisition by children, and language disorder in schizophrenic patients; goes through self-consciousness, sex, and the emergence of culture; and ends with a statistical analysis of the geographical and social distribution of economic power, discussing a host of involved factors, from cultural trends to intra- and international cooperation and conflict.

In the second chapter, *The Economy of Words*, Zipf revisited a finding which he had already advanced more that a decade earlier in his *The Psycho-Biology of Language* (Zipf, 1936), and that is now widely known as Zipf's law. Its original formulation establishes that, in a sizable sample of language —a text, or a piece of speech— the number of words $N(n)$ which occur exactly n times decays with n as

$$N(n) \sim n^{-\zeta} \tag{2.1}$$

for a wide range of values of n. While the exponent ζ varies from text to text, it is frequently found that $\zeta \approx 2$ (Zipf, 1936). Later on, Zipf favored an alternative, but equivalent, formulation (Zipf, 1949): if the words in the sample in question are ranked in decreasing order by their number of occurrences —with rank $r = 1$ for the most frequent word, rank $r = 2$ for the second most frequent word, and so on— the number of occurrences n is, to a good approximation, inversely proportional to a power of r:

$$n(r) \sim r^{-z}, \tag{2.2}$$

where, usually, $z \approx 1$. Sometimes, this relation is expressed in terms of the frequency of the word of rank r, $f(r)$, given by the ratio between $n(r)$ and the total length of the language sample. The relation between frequency and rank is the same as for the number of occurrences: $f(r) \sim r^{-z}$.

In his two books, Zipf profusely illustrated the *"hyperbolic"* relation between number of words, number of occurrences, and rank. Among others, he

provided examples from a collection of English texts taken from newspapers, compiled by R. C. Eldridge (1911) and containing more than 6000 different words; from James Joyce's *Ulysses*, with counting results by M. L. Hanley and M. Joos (Hanley, 1937); from four Latin plays by Plautus —*Aulularia*, *Mostellaria*, *Pseudolus*, and *Trinummus*— and from samples of Peiping Chinese dialect. These two latter instances showed that the power-law relations given by Eqs. (2.1) and (2.2) were valid beyond a single language, even holding across different linguistic families. Zipf's law thus emerged as a truly universal feature of human language.

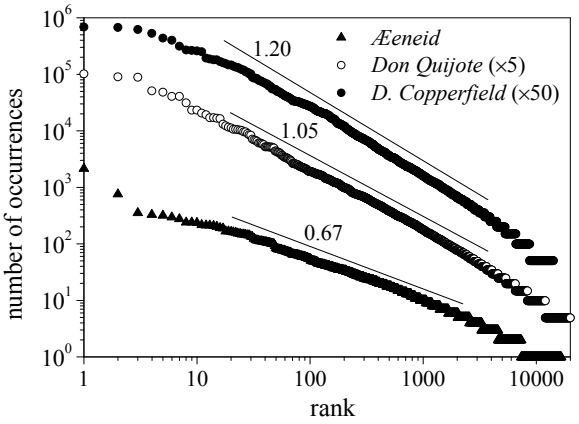

Figure 2.1: Number of occurrences as a function of the rank for the words of *Æneid* (in Latin), *Don Quijote* (in Spanish), and *David Copperfield* (in English). *Don Quijote*'s and *David Copperfield*'s data have been shifted upwards for clarity, multiplying the number of occurrences by factors of 5 and 50, respectively. Straight lines have the slopes of least-square fittings of each data set, calculated between ranks 50 and 1000. Slopes are indicated by labels. Adapted from Zanette and Manrubia (2007).

Today, Zipf's law has been verified for a multitude of texts in hundreds of languages, both extant and extinct. Even the mysterious Voynich manuscript, written in an unknown language using a not-yet-deciphered script, has been shown to comply with Eq. (2.2) (Landini, 2001). Figure 2.1 shows the number of occurrences n as a function of the rank r for three books in different languages: Virgil's *Æneid* (in Latin), Miguel de Cervantes's *Don Quijote* (in Spanish), and Charles Dickens's *David Copperfield* (in English). Their lengths (in number of words) are, respectively, $T = 63825$, 376629, and 363128, and their lexicon sizes (number of different words) are $V = 16656$, 23168, and 14078. In this log–log plot, a power-law relation such as Eq. (2.2) is represented by a straight line whose slope equals the exponent in the power law.

The three data sets exhibit the typical profile in this kind of frequency–

rank plot. For small ranks, there is a "shoulder" where n varies relatively slowly with r (note, however, the two most frequent words in *Æneid*). The power-law dependence develops for intermediate values of r, from several tenths to a few thousands. In the figure, this is emphasized by the straight lines, whose slopes have been obtained from linear least-square fittings of the log–log plot, for $50 < r < 10^3$. Note the rather large variation in the estimated values for the Zipf power-law exponent of Eq. (2.2), from $z = 0.67$ for *Æneid* to 1.20 for *David Copperfield*. Finally, for large r, the number of occurrences tends to decrease slightly faster. Table 2.1 lists the fifteen most frequent words of the three books, along with their number of occurrences.

r	*Æneid*		*Don Quijote*		*D. Copperfield*	
1	*et*	2146	*que*	20398	*the*	13763
2	*in*	761	*y*	17946	*I*	13472
3	*nec*	353	*de*	17935	*and*	12332
4	*per*	332	*la*	10225	*to*	10503
5	*ad*	321	*a*	9692	*of*	8748
6	*atque*	295	*en*	8077	*a*	7991
7	*non*	280	*el*	8066	*in*	6253
8	*cum*	240	*no*	6166	*that*	5388
9	*tum*	239	*los*	4697	*was*	5317
10	*quæ*	239	*se*	4644	*my*	5207
11	*est*	218	*con*	4154	*it*	5042
12	*nunc*	214	*por*	3838	*her*	3875
13	*aut*	214	*las*	3426	*you*	3730
14	*hæc*	212	*lo*	3414	*me*	3619
15	*iam*	203	*le*	3380	*he*	3610

Table 2.1: The fifteen most frequent words in *Æneid*, *Don Quijote*, and *David Copperfield*, and their number of occurrences.

The equivalence between Eqs. (2.1) and (2.2) can be readily ascertained by noting that the rank of a word with n occurrences, $r(n)$, equals the number of words with n or more occurrences.[1] Namely,

$$r(n) = \sum_{n'=n}^{\infty} N(n') \approx \int_{n}^{\infty} N(n')\, dn', \tag{2.3}$$

where $N(n)$ is the number of words which appear exactly n times. If $N(n)$

[1] Strictly speaking, this is true when there are no other words with the same number of occurrences as the word in question. When two or more words occur the same number of times, their relative ordering in Zipf's ranking is arbitrary. In this case, Eq. (2.3) defines the rank of the word which, among the group of those with the same number of occurrences, is chosen to be ranked first. Words with the same value of n are very frequent for small n, i.e. for large ranks. The "ladder steps" in the data sets at the lower–right corner of Fig. 2.1 correspond to those words.

satisfies Eq. (2.1), the relation between r and n is given by Eq. (2.2) with

$$z = \frac{1}{\zeta - 1}, \tag{2.4}$$

which yields $z = 1$ for $\zeta = 2$.

Zipf proposed a qualitative explanation of the inverse relation between the number of words and the number of occurrences by invoking the principle of least effort, on the basis of the balance between the "work" done by the two agents involved in a communication event —the speaker and the hearer. From the speaker's perspective, Zipf argued, the most economic vocabulary consists of a single word conveying all the desired meanings to be verbalized. The hearer, on the other hand, *"would be faced by the impossible task of determining the particular meaning to which the single word in a given situation might refer"* (Zipf, 1949). This conflict between the speaker's and the hearer's tendencies to respectively reduce and increase lexical diversification, is solved in the seek for efficient communication by developing a vocabulary where a few words are used very frequently, while most words occur just a few times.

Whereas this qualitative argument contented George Zipf as for an explanation of the empirical data, much more recently a mathematical model was put forward to quantify the process by which a vocabulary diversifies as communication evolves under the pressure of the principle of least effort on both speaker and hearer (Ferrer i Cancho and Solé, 2003). In the model, the process of communication implies the exchange of information about a collection of k objects (the "meanings"), $\{m_1, m_2, \ldots, m_k\}$, and uses a set of l signals (the "words"), $\{w_1, w_2, \ldots, w_l\}$. A binary $l \times k$ matrix $A = \{a_{ij}\}$ establishes the connection between words and meanings: if word w_i is used to refer to meaning m_j, then $a_{ij} = 1$; otherwise, $a_{ij} = 0$. Generally, it is allowed that more than one word refers to the same meaning, so that there may be several $a_{ij} = 1$ for the same value of j. The sum $\sigma_j = \sum_i a_{ij}$ is the number of synonyms referring to meaning m_j.

Let $p(w_i, m_j)$ be the joint probability that word w_i is used in the communication process when the information refers to m_j, and assume that all meanings are referred to with the same probability: $p(m_j) = k^{-1}$ for all j. The conditional probability of word w_i given the occurrence of meaning m_j is $p(w_i|m_j) = a_{ij}/\sigma_j$. Then, using Bayes's theorem (see Sec. 1.1), the total probability of w_i reads

$$p(w_i) = \sum_j p(w_i, m_j) = \sum_j p(m_j) p(w_i|m_j) = \frac{1}{k} \sum_j \frac{a_{ij}}{\sigma_j}. \tag{2.5}$$

The entropy associated with the probability distribution over words (Sec. 1.3),

$$H = -\sum_{i=1}^{l} p(w_i) \log_l p(w_i), \tag{2.6}$$

is a suitable definition for the speaker's communication effort. In fact, its minimum $H = 0$ is attained for a single-word vocabulary, where $p(w_i) = 1$ for that only word, and $p(w_i) = 0$ otherwise. When, on the other hand, $p(w_i) = 1/l$ for all i, the entropy is maximal: $H = 1$.

Upon hearing word w_i, the hearer must infer its meaning. The conditional probability for inferring meaning m_j is $p(m_j|w_i) = p(w_i, m_j)/p(w_i)$. In accordance with Eq. (2.6), the hearer's effort can be evaluated as the weighted sum of the entropies associated with the distribution $p(m_j|w_i)$ over all the words heard:

$$H' = -\sum_{i=1}^{l} p(w_i) \sum_{j=1}^{k} p(m_j|w_i) \log_k p(m_j|w_i). \tag{2.7}$$

Again, H' varies between zero and one, respectively corresponding to the limits of minimal and maximal hearer's communication effort. The entropy H' can be interpreted as a measure of the average "noise" (or indeterminacy) with which information reaches the hearer.

In the model, the total cost of communication is defined by the function

$$\Omega(\lambda) = \lambda H' + (1 - \lambda)H, \tag{2.8}$$

where $\lambda \in (0, 1)$ is a tunable parameter. Ferrer i Cancho and Solé (2003) performed numerical simulations for $k = l = 150$ where, at each step, a few elements of the matrix A were switched between zero and one or *vice versa*, and the change was accepted if the cost function $\Omega(\lambda)$ decreased. They expected that, if Zipf's hypothesis were valid, the probabilities $p(w_i)$ would converge to a distribution compatible with the inverse relation between frequency and rank for some intermediate value of λ. As a measure of communication accuracy, they also recorded the mutual information between the probability distributions of words and meanings, defined as

$$I(w, m) = \sum_{j=1}^{k} p(m_j) \sum_{i=1}^{l} p(w_i|m_j) \log_l p(w_i|m_j) - \sum_{i=1}^{l} p(w_i) \log_l p(w_i), \tag{2.9}$$

as well as the relative lexicon size, L, defined as the ratio between the number of effectively used words and the total number of available words, l.

Figure 2.2 is a schematic representation of the results of simulations, after a large number of iterations of the dynamical process of switching the elements of matrix A. The left panel shows the word–meaning mutual information $I(w, m)$ as a function of λ. Two distinct regimes are clearly identified, separated by a sharp transition at $\lambda^* \approx 0.41$. For $\lambda < \lambda^*$, there is practically no informational correlation between words and meanings, which is to say that communication fails. Accordingly, the relative lexicon size L —shown in the right panel of the figure— vanishes. For $\lambda > \lambda^*$, on the

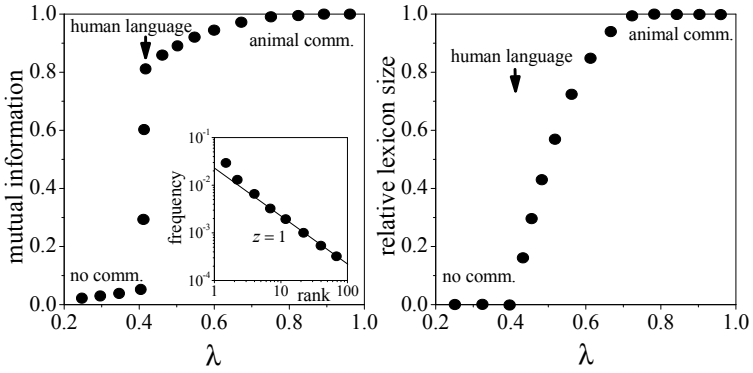

Figure 2.2: Schematic representation of numerical results from the quantitative model for the principle of least effort applied to the evolution of language. Left panel: The word–meaning mutual information $I(w, m)$ as a function of the tunable parameter λ, Eq. (2.8). Insert: Word frequency–rank relation for $\lambda = 0.41$. Right panel: Relative lexicon size L as a function of λ. Labels indicate the regimes of no communication and animal communication, and the transition at λ^*, which has been identified with human language. Adapted from Ferrer i Cancho and Solé (2003).

other hand, both $I(w, m)$ and L attain significant levels, and approach their maximal values for $\lambda \to 1$.

Remarkably, as shown in the insert of the left panel of Fig. 2.2, the analysis of the frequency–rank relation from the results of simulations satisfy Zipf's law with exponent $z \approx 1$ at the critical value λ^*, while the power-law relation breaks down for other values of λ. In the context of this model, therefore, human language appears to have been tuned by the principle of least effort at the edge of the transition between inviable and feasible communication.

Ferrer i Cancho and Solé (2003) conjectured that, during natural evolution, this tuning might have taken place in two broad stages. First, coinciding with the emergence of the anatomical and cognitive capabilities of ancestral animals to refer to an object by a signal, a transition occurred between the phase of no communication ($\lambda < \lambda^*$) to a phase of rudimentary referential signaling, with a high informational correlation between signals and objects ($\lambda > \lambda^*$). The phase of large λ, in fact, can be related to well-studied forms of communication in many animal species, which consist of relatively small signal repertoires with very specific meanings (Miller, 1981). Artificial languages, designed by humans for communication with and between machines, lie in the same phase, as they possess limited vocabularies and practically do not allow for synonymy.

As human social interactions grew in complexity, however, it became

necessary to drastically enlarge the potential number of signals, or words, used for communication. This, in turn, required a larger effort from the speaker which thus acquired a larger weight in the total cost of the process. The second stage, therefore, corresponded to a gradual decrease of the value of λ —which amounts to increasing the weight of the speaker's effort in the cost function of Eq. (2.8)— towards λ^*, where vocabularies can sensibly grow when the number of objects to be referred to by language increases.

In conclusion, Ferrer i Cancho and Solé's evolutionary model demonstrates that a convenient mathematical formulation of the principle of least effort does lead to Zipf's law, with $z \approx 1$. However, this result must be regarded cautiously. The model, in fact, describes the evolution of the frequencies of word usage in language as a whole. On the other hand, Zipf's law is known to be valid for the lexicons taken from single (or a small number of) texts. When many unrelated samples of the same language are combined into a single corpus, the resulting lexicon *does not* necessarily satisfy Zipf's law, as has been discussed by the same authors (Ferrer i Cancho and Solé, 2001a,b) and others (Kanter and Kessler, 1995; Montemurro, 2001; Montemurro and Zanette, 2002b).

In the next section, we discuss a mathematical model proposed in the 1950s by the economist and sociologist Herbert Simon —elaborating on previous ideas by the statistician Udny Yule— which focuses on the generation of a single text (Simon, 1955, 1957). While, quite artificially, the model conceives text production as a stochastic process, it is based on a few simple dynamical rules which make it suitable for explaining the appearance of algebraic relations of the kind of Zipf's law in many other phenomena. Properly interpreted, it provides a quantitative framework to understand the emergence of contextual elements as a text progresses.

2.2 Text Generation and Context Emergence

By the 1930s, when Zipf investigated the relation between frequency and rank in word usage, several examples were known of real-life statistical distributions with power-law dependence on their variables, of the type of Eqs. (2.1) and (2.2). Perhaps the most striking feature of these empirical data was the disparity of their origins. At the turn of the century, the economist Vilfredo Pareto had found that, in variety of economic systems and historical periods, the population whose income was above a given value u decreased in size as u^{-p} for large u, with $1 \lesssim p \lesssim 2$ (Pareto, 1896; Mandelbrot, 1997). City populations (Auerbach, 1913) and the size of firms (Gibrat, 1932) also display power-law tails in their distributions. Zipf (1949) illustrated these cases with his own examples, and added —among others— the number of service firms by kind in the U. S. A., 1939 (with barber shops, beauty parlors, and funeral companies ranking among the first

places). Willis and Yule (1922) reported on the power-law distribution of biological abundance, as determined by the number of species per genus in several plant and animal groups. Much later, this work was considerably expanded (Burlando, 1990, 1993), and extended to the taxonomic classification of human languages (Zanette, 2001). The size of populations bearing the same family name is now also known to be distributed according to a power law (Manrubia and Zanette, 2001). As an illustration of Zip's law in human groups, Fig. 2.3 shows the population as a function of the rank for the largest metropolitan areas, cities, and towns in three countries from different continents: India, Argentina, and France. In spite of the dissimilar historical, social, and economical circumstances of these countries, the slopes of their Zipf's rank plots are very close to each other.

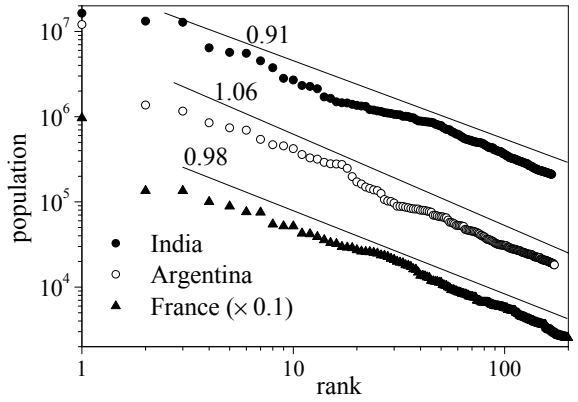

Figure 2.3: Population as a function of the rank for the largest urban settlements ($r \lesssim 200$) in India, Argentina, and France. The data for France have been shifted downwards for clarity, multiplying the population by a factor of 0.1. Straight lines have the slopes of least-square fittings of each data set, indicated by labels. Adapted from Zanette and Manrubia (2007).

Even the versatile imagination of Zipf's, when it came to apply the principle of least effort, was challenged by the diversity of phenomena that exhibit the same kind of mathematical relations as he had disclosed for the frequency of word usage. On the basis of a proposal by Yule (1925) for the case of biological abundance, Simon (1955) adopted a different perspective, assuming that very generic random mechanisms were at work in determining the statistical properties of such disparate systems. While Simon was well aware of the possibility of applying this idea to explain the distribution of, for instance, city populations or biological species per genus, he choose to introduce his model for the case of word frequencies. This is, precisely, the case that interests us here.

Consider the process of text generation as a sequence of events where

one word is added at each step, and let $N_t(n)$ be the number of different words that appear exactly n times when the text has reached a length of t words. Simon's model proposes the following two dynamical rules for the subsequent step:
(i) With constant probability α, the word added at step $t+1$ is a new one, which has not occurred in the first t steps. Namely, new words appear in the text at a constant rate α.
(ii) With probability $1-\alpha$, the word added at step $t+1$ is one of the words that have already occurred in the text. This recurrent word is chosen with a probability proportional to the number of its previous appearances.
Simon (1955) preferred to replace the second part of rule (ii) by a less stringent condition: the probability that the $(t+1)$-th word is a word that has already appeared exactly n times is proportional to $nN_t(n)$, i.e. to the total number of occurrences of all the words that have appeared exactly n times. However, as for the ensuing mathematical analysis, both formulations are equivalent.

Approximating the expectation value of the number of different words with exactly n occurrences at step $t+1$ by $N_{t+1}(n)$ itself, rules (i) and (ii) make it possible to write a recursive equation for $N_t(n)$, namely,

$$N_{t+1}(1) - N_t(1) = \alpha - \frac{1-\alpha}{t} N_t(1) \tag{2.10}$$

for $n = 1$, and

$$N_{t+1}(n) - N_t(n) = \frac{1-\alpha}{t} \left[(n-1)N_t(n-1) - nN_t(n)\right] \tag{2.11}$$

for all $n > 1$. The first term in the right-hand side of Eq. (2.10) represents the contribution to $N_{t+1}(1)$ of the word that appears for the first time at step $t+1$. Other terms in both equations are gain and loss contributions associated to the appearance of a word with, respectively, $n+1$ and n previous occurrences. In this formulation, Simon's model can be interpreted as a dynamical system for the function $N_t(n)$, where the running text length t plays the role of a discrete "time" variable. The above equations for $N_t(n)$ should be solved for a given "initial condition," $N_{t_0}(n)$, which stands for the distribution of occurrences of the words that have already been added to the text at the point t_0 at which the model's dynamical rules begin to act.

Equations (2.10) and (2.11) do not have a stationary solution, in the sense that an asymptotic, t-independent form for $N_t(n)$ does not exist. In fact, as the running text length grows, the normalization $t = \sum_n nN_t(n)$ must increase accordingly. A "steady-state" solution, however, can be sought for by assuming that, for large t, the number of different words with n occurrences satisfies $N_{t+1}(n)/N_t(n) = (t+1)/t$, for all n (Simon, 1955). This amounts to postulate the existence of a stationary profile $P(n)$ for $N_t(n)$

such that $N_t(n) = tP(n)$. Indeed, Eqs. (2.10) and (2.11) yield t-independent equations for $P(n)$, whose solution reads

$$P(n) = \frac{\alpha}{1-\alpha} B\left(n, \zeta\right). \tag{2.12}$$

Here, $B(n, \zeta)$ is the Beta function (Abramowitz and Stegun, 1972), and $\zeta = 1 + (1-\alpha)^{-1}$.

For small values of α ($\lesssim 0.1$) and for all $n \geq 1$, the above solution for the profile $P(n)$ is very well approximated by the power-law function

$$P(n) \approx \frac{\alpha}{1-\alpha} \Gamma(\zeta) n^{-\zeta}, \tag{2.13}$$

where $\Gamma(\zeta)$ is the Gamma function. This yields for $N_t(n)$ the form given by Zipf's law, Eq. (2.1) or, equivalently, Eq. (2.2) with $z = 1 - \alpha$. Since the probability of appearance of new words must necessarily be larger than zero, the exponent of the frequency–rank relation predicted by Simon's model is always lower than unity: $z < 1$. The characteristic value $z = 1$ is obtained in the limit $\alpha \to 0$, when the appearance of new words becomes extremely rare. In an actual text, this condition is expected to hold truer as the text progresses and becomes longer. While, in its original form, Simon's model is not able to explain Zipf's exponents z larger than one, extensions of the model that predict a wider range for z have been proposed (Zanette and Montemurro, 2005). They are reviewed in the next section.

Note that $N_t(n) = tP(n)$, with the profile $P(n)$ given by Eq. (2.12), is an *exact* solution to Simon's model equations (2.10) and (2.11). However, it does not represent their general solution, but a solution for a specific initial condition $N_{t_0}(n) = t_0 P(n)$ which already exhibits the profile $P(n)$. Due to the linearity of Eqs. (2.10) and (2.11), the general solution to Simon's model is a sum of the above special solution, $tP(n)$, plus a contribution from the initial condition. It can be seen that this latter contribution is modulated by the factor $t^{-(1-\alpha)}$ and, therefore, fades out as the text grows. At the same time, it shifts towards increasingly larger values of n so that, for long texts, its effects are appreciable only for very large n (Manrubia and Zanette, 2001). In the remaining range, below moderately large numbers of occurrences, $N_t(n)$ is dominated by the special solution with the power-law profile $P(n)$. Hence, Simon's model predicts that a power-law dependence between number of words, occurrences, and ranks should hold for small to moderately large values of n —or, in other words, for the lower ranks in Zipf's word lists (large r). For the higher ranks, on the other hand, deviations from Zipf's law are expected. These predictions are in good agreement with the frequency–rank relation found in real texts, as illustrated by Fig. 2.1. Only the faster decay of the number of occurrences for very large values of r is not accounted for by the model.

The random dynamics involved in Simon's model rules are a combination of additive and multiplicative stochastic processes for the number of occurrences n (Sornette, 2000). Rule (i) defines an additive process by which the number of words with $n = 1$ grows stochastically, at a constant average rate α. Rule (ii), which is the key ingredient in the emergence of Zipf's law from the model, describes in turn a stochastic reinforcement in the occurrence of words. Words which have already appeared a large number of times are more likely to be used again than those whose previous occurrences were rarer. This multiplicative process would give rise to an exponential explosion in the usage of single words, were it not by the fact that, at each step, there is a kind of competition for appearance between already used words among themselves and with those that may appear for the first time. As the result of this competition, in fact, the occurrences of any given word become on the average less and less frequent: in Simon's model, the average frequency of a specific word in a growing text of length t decays exponentially with the length, as $\exp(-\alpha t)$.

Equations (2.10) and (2.11) can be seen as the average evolution law deriving from a special case of a very general additive–multiplicative stochastic process for the number of occurrences of a word at step t, n_t, as governed by the following time-discrete Langevin equation (Sornette, 1998, 2000):

$$n_{t+1} - n_t = a_t + b_t n_t. \tag{2.14}$$

Here, a_t and b_t are random variables drawn, at each step, from suitably chosen probability distributions $f(a)$ and $g(b)$. The first and the second term in the right-hand side represent, respectively, the additive and the multiplicative contributions to the evolution. According to Eq. (2.14), the probability $p_t(n)$ that, at step t, the stochastic variable attains a given value n —which, in the context of our discussion, is proportional to the number of words with exactly n occurrences— can have a complicated analytical form, depending on the distributions from which a_t and b_t are chosen. Provided that $f(a) \neq 0$, it can however be proven that, for large t and within a wide range of values of n, the probability always behaves as (Sornette, 1998)

$$p_t(n) \sim n^{-1-\gamma}, \tag{2.15}$$

where the exponent γ is determined by the identity

$$1 = \int g(b)(1+b)^{\gamma} db. \tag{2.16}$$

This points out that power-law distributions as those implicit in Zipf's law, Eqs. (2.1) and (2.2), are inherent to generic additive–multiplicative stochastic processes, and are therefore expected to emerge in a large variety of disparate systems, as long as they are driven by random events (Sornette, 2000).

Although it is obvious that the process of creation of any meaningful text, in no matter which language, lies far from being a sequential random choice of words, the above results show that Simon's model provides a stochastic mimic that correctly reproduces some salient features in the statistics of word usage. Zipf's law is one of these features. The model's rules have been interpreted as a representation of the basic mechanisms by which context emerges as a text is generated (Montemurro and Zanette, 2002b; Zanette and Manrubia, 2007). Context is the global property of a structured message that sustains its coherence and intelligibility (van Eemeren, 2001). A long chain of words, even if they constitute a grammatically correct language sample, would result incomprehensible if it does not succeed at defining a contextual framework. It is in this framework —created by the message itself as genre, style, form, subject, tense, person, etc. are introduced and developed— that its perceptual elements become integrated into a meaningful, self-consistent structure.

Context emerges from the mutually interacting meanings of words, and represents a collective expression of the semantic contents of the message, arising from the multiple structured relations between language elements. As words are successively added to the text, a context is built up which favors the later appearance of some words, in particular, those that have already been used, and inhibits the use of others. Reinforcement of the contextual constructions by repetition of perceptual elements is one of the basic ingredients in the conception of intelligible structures, and in our brain's response to their reception, including the creation and retrieval of memories (Brown and Hagoort, 2000). This notion lies at the basis of the cognitive processes related to human communication through language.

This connection between context emergence and reinforcement of word occurrence can be tested in other forms of communication, conveying information whose nature is different from that of language, but where the notion of context —as the framework that sustains the intelligibility of a piece of information— is still meaningful. One example is provided by music. As a human universal related to communication, the acquisition, generation, and perception of music share at least some basic neural mechanisms with language (Maess *et al.*, 2001; Patel, 2008). The appealing affinity between the cognitive processes triggered by music and language has naturally led to the attempt of extending concepts and methods of linguistics to the domain of musical expression (Bernstein, 1973; Lerdahl and Jackendoff, 1983). In contrast to language, however, music lacks functional semantics. Generally, the musical message does not convey information about the extra-musical world. A conventional correspondence between musical elements and non-musical objects or concept is therefore irrelevant to the cognitive functions of music. Assigning extra-musical meaning to a musical message is basically an idiosyncratic matter, and cannot be expected to yield universal results.

Context, on the other hand, is a concept as essential to music as it is to

language (Zanette, 2006, 2008). As discussed above for a grammatically correct array of words, a sequence of music events turns out to be unintelligible if it does not create a musically meaningful structure. In Western music, context is determined by a hierarchy of intermingled patterns occurring at different time scales. For the occasional listener, the most evident contribution to musical context comes from the melodic material, whose repetitions and variations shape the thematic base of a composition. Tonal and rhythmic features of melody phrases constitute the substance of musical context at that level. At larger scales, the recurrence of long sections and certain standard harmonic progressions determine the musical form and, frequently, defines style. Crossed references between different movements of a given work establish correlations over even longer scales. At the opposite end of time scales, a few notes are often enough to determine tempo, rhythm, and tonality, through their relative durations and pitches (Krumhansl, 1990).

As for the "building blocks" of musical context —which replace words in collectively yielding coherence and comprehensibility to a musical message— several choices are possible among the perceptual elements that make the message intelligible. Zipf (1949) showed that a power-law relation holds between the number of pitch intervals between contiguous notes and their size in all movements of a bassoon concerto by Mozart. Other authors studied Zipf's law for the distribution of single notes, melodic digrams and trigrams, and interval doublets and triplets (Boroda and Polikarpov, 1988; Manaris *et al.*, 2003; Zanette, 2006). Single notes endowed with pitch and duration are particularly transparent as for their role in the construction of context, determining the basis of tonality and rhythm. Figure 2.4 displays Zipf's plots for the number of occurrences as a function of the rank for single musical notes in four Western compositions (Zanette, 2006). Although they do not exhibit a range of power-law dependence, their mutual similarities are striking.

Curves in the plots of Fig. 2.4 are fittings obtained from an extension of Simon's model where, first, an upper limit is imposed to the total number of occurrences of individual notes and, second, the rate of appearance of new notes is not a constant but varies with the running length t as $t^{\nu-1}$. This latter generalization is discussed in the next section. The multiplicative stochastic process involved in Simon's model, on the other hand, acts as specified by the model's rule (ii). Under these conditions, it can be seen that the number of occurrences n varies with the rank r as

$$n \sim (a + br)^{-1/\nu}, \tag{2.17}$$

where a and b are constants. Because the compositions under study are relatively short sequences, the values of r are never very large and the power-law tail of Eq. (2.17) does not develop. The quality of the fittings is however very good, supporting the view that a multiplicative process also underlies the establishment of frequency–rank relations in music.

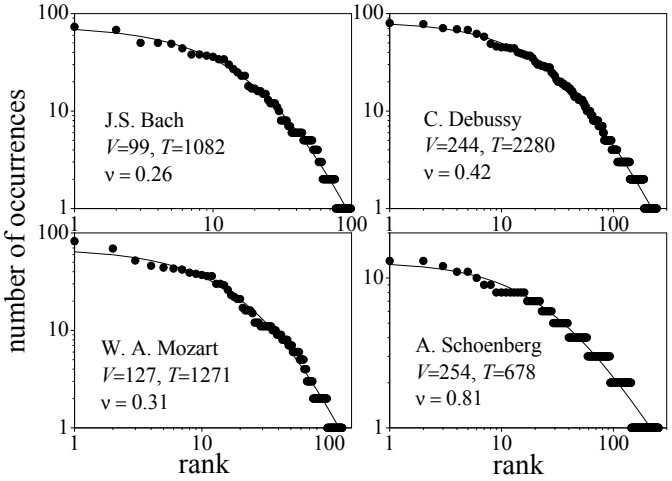

Figure 2.4: Number of occurrences of musical notes (endowed with pitch and duration) n, as a function of their rank r, for four musical compositions: J. S. Bach's *Prelude in D minor*, from *The Well Tempered Clavier, Book II*; W. A. Mozart's *Piano sonata in C*, K545, first movement; C. Debussy *Suite Bergamasque*, second movement; and A. Schönberg's *Drei Klavierstücke*, Op. 11, N. 1. Curves are fittings with a function of the form $n \sim (a+br)^{-1/\nu}$ (see text for details). Labels indicate the length of each piece T, in number of notes, the number of different notes V, and the exponent ν. Adapted from Zanette (2006).

Note, finally, the rather large variations in the exponent ν between different composers. The relatively small values of ν for Bach and Mozart are an indication of a compact "lexicon," determining a robust context that remains stable as que musical sequence progresses. On the other hand, the large exponent of Schönberg's composition reveals an abundant "lexicon," defining a ductile, less steady context, typical of its atonal style.

2.3 Extensions of Simon's Model: Heaps's Law and Human Populations

As we have shown in the preceding section, Simon's model establishes a relation between the exponent z in Zipf's frequency–rank power law, Eq. (2.2), and the average rate of appearance of new words in a text, α. This average rate, which gives the probability that the word at any given place in the text occurs there for the first time, is the only parameter in the model. The relation $z = 1 - \alpha$ implies that, according to Simon's prediction, the exponent in Zipf's law should be lower than one. In contrast to this prediction, however, it is not difficult to find language samples where the best fitting of

the frequency–rank relation yields $z > 1$. Figure 2.1 shows that *Don Quijote* and *David Copperfield* are two such counterexamples.

While it is arguably the simplest hypothesis regarding the appearance of new words in a stochastic model for the generation of a text, the assumption that the rate α is constant all along the process is unrealistic. Figure 2.5 shows the number of different words as the text progresses, as a function of the running text length t, for the three books considered in Fig. 2.1. If the rate at which new words appear were a constant, the number of different words would grow linearly with t, and the data sets in this log–log plot would be represented by straight lines with slope equal to one. In contrast, the plot shows that the growth is sub-linear, being acceptably approximated by a power of t for large values of the running text length.

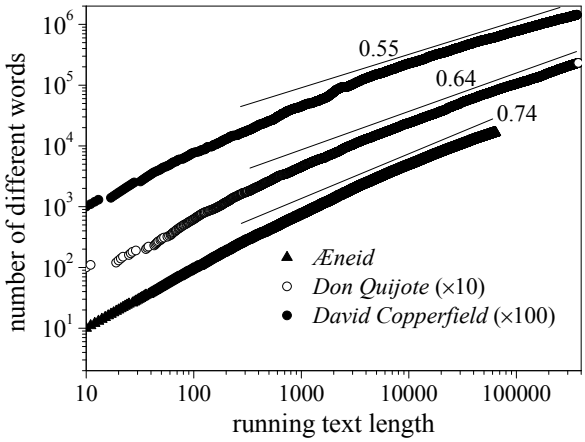

Figure 2.5: Number of different words as a function of the running text length t for *Don Quijote*, *David Copperfield*, and *Æneid*. *Don Quijote*'s and *David Copperfield*'s data have been shifted upwards for clarity, multiplying the number of occurrences by factors of 10 and 100, respectively. Straight lines have the slopes of least–square fittings of each data set, calculated in the range of large t. Slopes are indicated by labels.

The power-law relation between the number V of different words in a text and the text length T, $V \sim T^{\nu}$, is usually referred to as Heaps's law (Heaps, 1978). The sub-linear growth of V with T is insured if the (positive) Heaps exponent ν is lower than one. In principle, Heaps's law stands for the dependence of the total number of different words on the total length of a given text. It is however customary to extend the law to the way in which the number of different words grows as a text progresses or, more generally, as increasingly large portions of a given language corpus are considered (Gelbukh and Sidorov, 2001; Lü *et al.*, 2010). Within this interpretation, Heaps's law establishes that the rate of appearance of new words decays

with the running text length t as

$$\alpha(t) = \alpha_0 t^{\nu-1}, \tag{2.18}$$

where $\alpha_0 < 1$ is a constant. The slopes of the data sets in Fig. 2.5, indicated by labels, are direct estimations of the Heaps exponent ν.

Note, comparing Figs. 2.1 and 2.5, the correlation between the frequency–rank exponent z in Zipf's law and the Heaps exponent ν for the three books considered there. Large values of z correspond to small ν, and *vice versa*. This inverse relation can be qualitatively understood by noting that texts written in highly inflected languages —like Latin, where many different words derive from the same root, through verb conjugation and noun declension— are expected to have richer lexicons (in number of different words) than texts written, for instance, in English, where a single verb form is used for several persons and tenses, a single noun for many declension cases, and so on (Zanette and Montemurro, 2005). This difference is dramatically illustrated by comparison of the lexicon sizes and total lengths of *Æneid* and *David Copperfield*: while the lexicon of the Latin poem is larger than that of the English novel by some 15%, the latter work is almost six times longer than the former. Spanish, which has rich inflections for verb conjugation but does not use declensions, is in this sense an intermediate case between Latin and English.

As a Latin text progresses, it is expected that the number of different words grows relatively faster than in English, thus exhibiting a larger value of the Heaps exponent ν. Once the Latin text is finished, in turn, the total number of word occurrences —namely, the total text length— is distributed among a larger number of different words and, therefore, the frequency–rank distribution is expected to display a flatter profile. Hence, the Zipf exponent z should be smaller. As shown in the following, Simon's model is able to explain this inverse relation between ν and z, if generalized to admit that the probability of occurrence of new words can vary along the text. Moreover, it predicts that the Zipf exponent z is larger than one, at difference with the original version of the model.

A convenient approximate way to deal with Simon's model under the assumption that the probability of occurrence of new words varies as the text progresses, is to conceive the two variables t and n in Eqs. (2.10) and (2.11) as continuous quantities (Zanette and Montemurro, 2005). Denoting $N_t(n) \equiv N(n,t)$, we have

$$\dot{N}(1,t) = \alpha(t) - \frac{1-\alpha(t)}{t} N(1,t) \tag{2.19}$$

for $n = 1$, and

$$\partial_t N(n,t) = -\frac{1-\alpha(t)}{t} \partial_n [nN(n,t)] \tag{2.20}$$

for $n > 1$. Once the first of these equations is solved,

$$N(1,t) = N(1,t_0)\epsilon(t) + \epsilon(t)\int_{t_0}^{t} \frac{\alpha(t')}{\epsilon(t')}dt' \tag{2.21}$$

with

$$\epsilon(t) = \exp\left[-\int_{t_0}^{t} \frac{1-\alpha(t')}{t'}dt'\right], \tag{2.22}$$

its solution serves as a boundary condition for the second, in the limit $n \to 1$.

Let us now assume that the probability of occurrence of new words is given by Heaps's law, as in Eq. (2.18). Since $\alpha(t)$ decays following a power law, we have $1 - \alpha(t) \approx 1$ for sufficiently large values of t. In this limit, the dominant contribution to the solution to Eq. (2.19) is a growing power of t: $N(1,t) \approx \alpha_0 t^{\nu}/(1+\nu)$. Within the same approximation, in turn, the solution to Eq. (2.20) has the form $N(n,t) \approx n^{-1}H(n/t)$, where $H(x)$ is an arbitrary function. Using the boundary condition for $n \to 1$ yields

$$N(n,t) \approx \frac{\alpha_0}{1+\nu} t^{\nu} n^{-1-\nu}. \tag{2.23}$$

Comparison with Eqs. (2.1) and (2.2) shows that, within this extension of Simon's model, the Zipf exponents are $\zeta = 1+\nu$ and $z = \nu^{-1}$. The frequency–rank Zipf exponent z is, consequently, larger than one and exhibits a simple inverse relation with the Heaps exponent ν. While substantial quantitative differences remain between this prediction for the relation of z and ν and the actual values of the exponents, the result is in qualitative agreement with the empirical data for the three books considered above.

Using a more heuristic approach, which does not involve a dynamical model for text generation, Lü *et al.* (2010) have found a similar relation between the Zipf and Heaps exponents. Their approach was satisfactorily applied to books in several European languages of the Romance and Germanic families, as well as to a corpus of keywords from scientific journals in English. A deep understanding of the connection between Zipf's and Heaps's law, though, still requires extensive empirical work and quantitatively robust models.

Turning now back the attention to the original form of Simon's model, Herbert Simon himself pointed out that the same stochastic dynamical rules may be useful to portray many of the phenomena which display power-law distributions in their statistical properties (Simon, 1955, 1957). From an abstract perspective, the model describes the random growth in size of certain object classes, with a growth rate proportional to the class size itself. In the case of text generation, each class corresponds to a different word, and the class size is the number of its occurrences in the text. This stochastic multiplicative growth is added with a random process by which new classes, with the minimal size, are created at a fixed rate.

Stochastic multiplicative growth is the basic dynamical process that governs the size of biological populations —and, among them, human communities. Populations of living beings are subject to a multitude of actions of disparate origins, involving complex interplay between endogenous and external factors related to the interaction with the ecosystem and the physical environment. The fluctuating nature of these factors imply that the parameters that drive the evolution of the population —such as the birth and death rates— change with time in irregular ways. On the average, however, if births are consistently more frequent that deaths, a population will exponentially grow in size until it reaches the corresponding carrying capacity of the ecosystem (Murray, 2002).

Additionally, due to factors of social, economic, and cultural origin, human communities spontaneously split into classes or groups of different identity, whose sizes evolve to a large extent independently from each other. Three paradigmatic examples are given by the populations of urban settlements (see Fig. 2.3), the speakers of different languages, and the groups of people that bear a common family name —with each class represented, respectively, by a settlement, a language, and a family name.

In the latter instance, the rules that govern the transmission of a family name from parent to child insure, first, that the growth of the class size follows the same stochastic multiplicative law as the population itself. Second, since the exchange of individuals between classes is negligible, each class preserves its identity along very long times. Finally, it is expected that new family names are created from time to time, essentially due to changes is pronunciation and spelling (Manrubia *et al.*, 2003). While all these features are in agreement with the mechanisms inherent to Simon's model, there is a crucial difference between the growth of a family and the increasing occurrences of a word in a text: in contrast to words, human beings die, and family names eventually disappear. Hence, to apply Simon's model to human populations, mortality needs to be taken into account.

The extension of Simon's model to include death events has been discussed, precisely, in the context of the distribution of family names (Manrubia and Zanette, 2001, 2002). It consists of considering evolution steps —analogous to the steps at which words are added to a text— each of them corresponding to the birth of a new person, endowed with a family name chosen from the preexisting population with a probability proportional to its own frequency. With probability α, however, the newborn adopts a new family name, not previously used by the population. At the same step, a randomly chosen individual dies with probability μ. The generalization of Eqs. (2.10) and (2.11) to this situation is rather straightforward, though finding its solution turns out to be more cumbersome. A suitable approximation of a continuous version of the equations, of the type of Eqs. (2.19) and (2.20), yields for the number of family names borne by exactly n people

at large t (Manrubia and Zanette, 2002):

$$N(n,t) \sim \frac{P(t)}{n} U\left(\frac{1-\mu}{1-\alpha-\mu}, 0, 2\frac{1-\alpha-\mu}{1-\alpha+\mu} n\right) \tag{2.24}$$

with $P(t)$ the total population at step t, and $U(a,b,x)$ the logarithmic Kummer's function (Abramowitz and Stegun, 1972). In a wide range of values of n, this solution behaves as a power law,

$$N(n,t) \sim n^{-1-\frac{1-\mu}{1-\alpha-\mu}}, \tag{2.25}$$

thus generalizing the frequency–rank exponent to

$$z = 1 - \frac{\alpha}{1-\mu}. \tag{2.26}$$

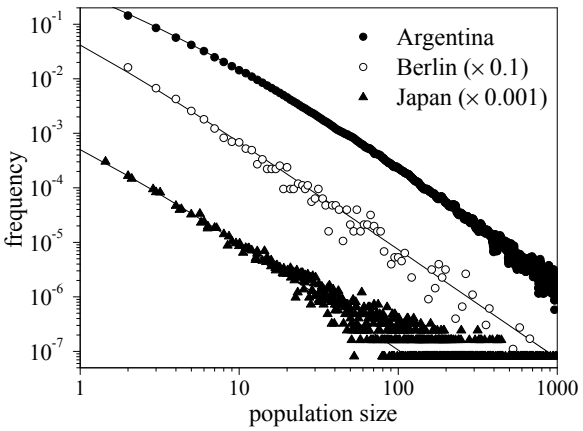

Figure 2.6: Normalized frequency of populations bearing the same family name, as a function of the population size, for two countries and a city. Argentina: almost 350000 family names from the whole 1996 telephone book; Berlin: almost 6400 family names beginning by A in the 1996 telephone book; Japan: taken from Miyazima *et al.* (2000). The data for Berlin and Japan have been shifted downwards for clarity, multiplying the frequency by factors of 0.1 and 0.001, respectively. Curves stand for nonlinear fittings with the functional dependence given by Eq. (2.24). Adapted from Manrubia and Zanette (2002).

Figure 2.6 shows empirical data for the frequency of populations with the same family name as the function of their size for two countries and a city. Data sets were collected by counting entries in phone books. The curves correspond to fittings with the functional dependence predicted by Eq. (2.24). Overall, the agreement between the data and the extended Simon's model discussed here is excellent. Only for large populations do

the Japanese data show a substantial difference with the prediction. This difference has been attributed to an effect of initial conditions (Manrubia and Zanette, 2002). While the distributions of family names now settled in Berlin and in Argentina have been evolving since the Middle Ages, when they began to be extensively used in Europe, present family names in Japan date from after the end of the nineteenth century (Miyazima *et al.*, 2000). A steady distribution in Japan is not yet established, and the difference is more apparent for large populations which, arguably, bear relatively older family names.

Finally, it is worthwhile pointing out that Simon's model, when extended to encompass death events, is equivalent to a branching process (Harris, 1963). The theory of branching processes has become the standard framework for the statistical description of the distribution of family names (Panaretos, 1989; Consul, 1991; Islam, 1995).

2.4 Is Zipf's Law Really Relevant to Language?

In his introduction to the 1965 edition of Zipf's *The Psycho-Biology of Language*, published fifteen years after the author's death, the Harvard psychologist George A. Miller offers to the reader a summary on Zipf's life and on his contributions to philology. Miller does not hesitate to remark the wealth of empirical evidence gathered by Zipf to support power-law relations in the frequency of word usage in several languages. About the origins of Zipf's law, however, he is definitely sceptical: *"Faced with this massive statistical regularity, you have two alternatives. Either you can assume that it reflects some universal property of the human mind, or you can assume that it represents some necessary consequence of the laws of probability. Zipf chose the synthetic hypothesis and searched for a principle of least effort that would explain the apparent equilibrium between uniformity and diversity in our use of words [...] Now,[...] Suppose that we acquired a dozen monkeys and chained them to typewriters until they had produced some very long and random sequence of characters. Suppose further that we define a 'word' in this monkey-text as any sequence of letters occurring between successive spaces. And suppose finally that we counted the occurrences of these 'words' in just the way Zipf and others counted the occurrences of real words in meaningful texts. When we plot our results in the same manner, we will find exactly the same 'Zipf curves' for the monkeys as for the human authors. Since we are not likely to argue that the poor monkeys were searching for some equilibrium between uniformity and diversity in expressing their ideas, such explanation seems equally inappropriate for human authors [...] So Zipf was wrong"* (Miller, 1965).

This rather drastic conclusion of Miller's was based on a remark by the mathematician Benoit Mandelbrot, first published in Mandelbrot (1951) and

later reproduced in several equivalent forms (Mandelbrot, 1997; Li, 1992). Mandelbrot proposed to let typewriting monkeys[2] use an alphabet of $M+1$ letters, including the blank space, to produce a random text in which the blank space appears with probability p_0, and each of the other M letters with probability $(1-p_0)/M$. Considering that any array of these M letters between two consecutive blank spaces is a "word," there are exactly M^l different words of length l (i.e. formed by l letters), and each of them occurs with exactly the same frequency

$$f(l) = p_0 \left(\frac{1-p_0}{M} \right)^l . \tag{2.27}$$

Since the probability of any given word decays exponentially as its length increases, the rank of a word of length l used in a long "monkey-text" will be of the order of the total number of different words shorter than l:

$$r(l) \sim \sum_{i=1}^{l-1} M^i \sim M^l . \tag{2.28}$$

Elimination of the word-length l from Eqs. (2.27) and (2.28) yields a power-law frequency–rank relation $f(r) \sim r^{-z}$ with

$$z = 1 + \frac{\ln(1-p_0)}{\ln M} < 1. \tag{2.29}$$

According to Mandelbrot's model, then, a random text satisfies Zipf's law with a frequency–rank exponent $z < 1$, which approaches the characteristic value $z = 1$ for large M and/or small p_0. Modern European languages have $M \approx 25$ and $p_0 \approx 0.2$, which gives $z \approx 0.93$. More generic (Markovian) random processes have also been shown, by Mandelbrot and others, to lead to Zipf's law (Mandelbrot, 1955; Kanter and Kessler, 1995).

Because of its extreme simplicity, Mandelbrot's model stands as a kind of null hypothesis for Zipf's law. If it could not be disproved, the conclusion would be that, as for the frequency of word usage, it would impossible to discern between a sample of real language and a random sequence of letters and blank spaces. In other words, Zipf's law would contain no relevant information about the statistics of language. At the same time, a model

[2]Mandelbrot, however, was not the first to invoke typewriting monkeys to generate a random text. In his 1913 article *Mécanique Statistique et Irréversibilité* and his 1914 book *Le Hasard*, the statistician Émile Borel asserted that it was less likely that the laws of statistical mechanics would ever be violated than a million monkeys typing ten hours a day would produce in their lifetime all the books of the richest libraries of the world. More than a decade later, the physicist Arthur Eddington, in his *The Nature of the Physical World* wrote: *"If an army of monkeys were strumming on typewriters they might write all the books in the British Museum. The chance of their doing so is decidedly more favourable than the chance of the molecules returning to one half of the vessel [where they were initially confined]"*.

based on linguistically more sensible hypotheses, such as Simon's model, would stand as a unnecessarily sophisticated explanation for a trivial feature of symbol sequences. As Mandelbrot himself put it, there would be *"nothing more to Zipf's law"* with $z < 1$ (Mandelbrot, 1997). This remark led in fact to a lively academic discussion between Mandelbrot and Simon, along a series of six papers starting with Mandelbrot (1959) and ending with Simon (1961).

While Mandelbrot's random-text model is still sometimes invoked as a proof that Zipf's law is devoid of any linguistically relevant meaning, it is not difficult to show that the model makes some subsidiary predictions which are in fact *not* met by real language samples. Perhaps the most obvious is that, according to Eq. (2.29), the Zipf frequency–rank exponent should depend on the alphabet size of each language. It is true that the model unrealistically assumes that all (non-space) letters are used with the same probability, but the same dependence with the number of different letters is expected to hold also for more general probability distributions of letter occurrence. Now, such dependence is not supported by empirical data, even when the alphabets of some languages where Zipf's law has been amply verified differ by some 20% in size —for instance, English with 26 letters versus Latin with 21 letters. According to Eq. (2.29), an English text's frequency–rank relation would change if it were translated word by word into Latin. Differences in the Zipf exponents between languages, on the contrary, seem rather to be related to the diverse use of inflections, as discussed in the preceding section.

Also, implicit in Mandelbrot's model is a very specific prediction on the distribution of word lengths, in number of letters, as given by Eq. (2.27): the probability that a word chosen at random from a text has l letters should decay exponentially with l. Figure 2.7 shows the distribution of word lengths in *Æneid*, *Don Quijote*, and *David Copperfield*, with each book written in its original language. It is apparent that these empirical distributions differ substantially from the exponentials predicted by the model. The straight line in the log–linear plot of the figure stands for Mandelbrot's prediction with $M = 26$ (English alphabet) and for $p_0 = 0.246$, which coincides with the frequency of blank spaces in *David Copperfield*. In the empirical distributions, the maximal frequencies stand for words with two to five letters. This responds to a balance between the trend to make words as short as possible, and the variety of sounds that can be achieved by combining more than just a few letters. Zipf (1949) discussed this trade-off, not unexpectedly, in terms of the principle of least effort.

It might be argued, however, that the random-text model could be modified to account for a more realistic distribution of word lengths while still preserving the prediction of a power-law frequency–rank relation. Although it is difficult to provide a complete answer proving or disproving such conjecture, an example can be given that shows that changing the hypotheses

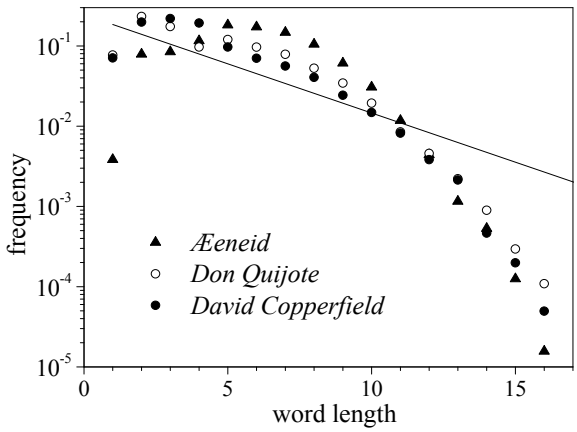

Figure 2.7: Distribution of word lengths for *Æneid*, *Don Quijote*, and *David Copperfield*, in log–linear scales. The straight line is the prediction of Mandelbrot's model, for a blank-space probability that equals the frequency of blank spaces in the sequence of letters of *David Copperfield*.

that lead to the exponential distribution of word lengths does modify Zipf's law. Suppose that a text is generated by choosing the sequence of letters at random, as in Mandelbrot's model, but controlling the appearance of blank spaces in such a way that all words have exactly the same length l. In this situation, all possible words have the same probability, so that in a long (but finite) text differences in the number of occurrences of different words will be governed by fluctuations. Taking into account that the number of different words is M^l, the number $N(n)$ of words with n occurrences in a text of length T will be proportional to a Gaussian distribution centered at $\langle n \rangle = T/M^l$ with mean square dispersion $\sigma_n \sim \sqrt{T/M^l}$. Replacing this form of $N(n)$ in Eq. (2.3) gives for the rank dependence of the number of occurrences an inverse error-function relation (Abramowitz and Stegun, 1972), which bears no connection with Zipf's law, Eq. (2.2).

Finally, Ferrer i Cancho and Elvevåg (2010) have put forward what is arguably the most important objection to Mandelbrot's model for Zipf's law: due to strong fluctuations originated in the random mechanisms at work, the frequency–rank relations generated by the model are much more irregular than those observed in real texts. These authors have quantitatively characterized such differences by means of statistical tests, for realizations of three variants of Mandelbrot's model. In the first version, the characters of the alphabet and the blank space were used all with identical probabilities. In the second, all characters except the space had the same probabilities, and the probability of the space was taken from real language samples. In the third, each character was allowed a different probability, taken again from real texts. The probabilities for different characters were taken mostly from

English books written during the nineteenth century, such as Dickens's *A Christmas Carroll* and Darwin's *On the Origin of Species*.

Not unexpectedly, the frequency–rank relations of random texts generated with the third variant of the model were closer to those of the real texts than for the other variants. However, the differences with real frequency–rank Zipf plots remained statistically significant. Ferrer i Cancho and Elvevåg concluded that, at the time of their work, a good fit of random texts to real Zipf's rank distributions had not yet been performed, so that the relevance of Zipf's law to language was not disproved. This conclusion warns on mistaking asymptotic results such as Mandelbrot's with actual realizations of random processes. At the same time, it leaves open the way to further research on Zipf's law based on linguistically compelling arguments.

Chapter 3

Long-Range Organization in Language Streams

Regularities in the distribution of frequencies of word usage, described by Zipf's law and analyzed in the preceding chapter, reveal an elementary but nontrivial level of organization in human language. However, the structural properties of linguistic patterns are obviously not exhausted by the power-law distribution of word frequencies. If the sequence of words in a real text were shuffled at random, completely destroying its original order, the most likely outcome would be an unintelligible array of disconnected words, with neither syntactic coherence nor meaning. Still, the number of occurrences of each different word would be the same in the shuffled text as in its original version. It is thus clear that, beyond word frequencies, grammatical constraints and semantic development define further contributions to the organization of language. The rules of syntax, which strongly vary between different human tongues, impose conditions on word ordering over scales that rarely surpass the scope of individual clauses or sentences. On the other hand, the unfolding of consistent semantic contents along the discourse may reach spans comprising from a few sentences to several paragraphs. In contrast with grammatical rules, moreover, the process of semantic construction —which is more directly related to the overall function of language as a communication system— is expected to be less dependent on the specific tongue where it takes place. It is over these long-range scales that a quantitative statistical approach to the description of linguistic structures turns out to be most fruitful.

This chapter addresses the study of statistical regularities in the ordered sequences that constitute written texts, be they seen as arrays of letters or words. In view of the available collection of statistical tools for the analysis of temporal signals and symbolic sequences, it becomes useful —from the perspective of quantitative linguistics— to think of language streams as the output of a dynamical system, perhaps governed by stochastic rules,

and susceptible of standard statistical analysis (see also Sec. 1.2). We first review the representation of a text, considered as a sequence of letters, by means of a random walk. The calculation of the diffusion exponents of this process discloses its anomalous transport properties, associated with long-range positive correlations in the successive appearance of letters along the text. A similar study, where —more sensibly from the linguistic viewpoint— texts are viewed as sequences of words instead of letters, shows that such correlations can be identified with fractal properties in word ordering. In both cases, comparison with shuffled texts where the internal structure of sentences is preserved, shows that long-range correlations are induced by organization across sentences, thus transcending the scope of grammatical rules. Finally, we discuss the highly heterogeneous distribution of certain frequent words along texts. This property, known as burstiness, is closely related to the emergence of long-range correlations. Its calculation has been proposed as a tool for the automated identification of keywords.

3.1 Letter Sequences and Random Walks

The transformation of a sequence of symbols into a random walk was put forward to study long-range organization in DNA sequences (Peng *et al.*, 1992), and was used to disclose their power-law correlations and their "patchy" or mosaic-like structure (Peng *et al.*, 1994; Stanley *et al.*, 1994). The first application to written texts also revealed the existence of power-law correlations (Schenkel *et al.*, 1993). Ebeling and Neiman (1995) proposed a variant in which texts were seen as sequences composed from a collection of 32 symbols, including the 26 letters of the English alphabet, the blank space, the period, the comma, the opening and closing parentheses, and the numeral symbol (#). Their method proceeded as follows. For each symbol k a binary string was constructed in such a way that its t-th element equaled 1 if k occurred at place t in the text, and 0 otherwise. The number of elements in the string was therefore equal to the length of the text, measured in number of symbols. For instance, for the letter *a* and for the blank space, Herman Meville's *Moby Dick* —which begins by *"Call me Ishmael..."*— was respectively represented by the strings 010000000000100... and 000010010000000....

The 32 strings thus constructed were used to generate a random walk in a 32-dimensional space. Starting at an arbitrary point in this space, at each time step i, the walker moved one step forward in the k-th direction if the i-th element in the corresponding string was 1. After a certain total time t, the walker had moved a total distance $y_k(t)$ along the k-th direction. Clearly, $y_k(t)$ was equal to the number of occurrences of symbol k until the t-th place in the text. The quantity

$$\sigma_k^2(t) = \langle y_k^2(t) \rangle - \langle y_k(t) \rangle^2 \tag{3.1}$$

measures the mean variance of the walker's displacement in the k-th direction. The brackets $\langle\cdot\rangle$ represent averages over all initial positions in the text. Over the whole 32-dimensional space the random-walk mean variance is given by $\sigma^2(t) = \sum_k \sigma_k^2(t)$.

If the random walker's motion is power-law correlated, it is expected that the mean variance of its displacement scales with time as

$$\sigma^2(t) \sim t^\alpha \tag{3.2}$$

for large t (Applebaum, 2005). Normal and anomalous diffusive behavior produce, respectively, $\alpha = 1$ and $\alpha > 1$. The latter is indicative of long-range correlations: in this case, mean square displacements are asymptotically larger than for uncorrelated random walks.

Ebeling and Neiman (1995) performed this analysis for a German edition of *The Bible*, the Brothers Grimm's *Tales*, and Meville's *Moby Dick*. Their lengths in number of symbols of the 32-symbol "alphabet" were, respectively, slightly above 4×10^6, 1.4×10^6, and 1.1×10^6. The upper curve in Fig. 3.1 represents the measurements of $\sigma^2(t)$ as a function of t for the German *Bible*. In this log–log plot, the well defined power-law dependence of the mean displacement variance is put in evidence by the linear profile of the data. The power-law exponent, given by its slope, is $\alpha \approx 1.698$. It is interesting to mention that, for an English version of the first half of the *Old Testament*, Schenkel *et al.* (1993) had obtained results corresponding to $\alpha \approx 1.74$. On the other hand, the values for Grimm's *Tales* and *Moby Dick* were sensibly lower, but still clearly above the uncorrelated case: $\alpha \approx 1.175$ and 1.241, respectively.

These results were then compared with shuffled versions of the sequence of symbols given by each text. Shuffling was applied at three levels: a random reordering of all symbols in the sequence, a random reordering of words (defined as the subsequences between successive spaces), and a random reordering of sentences (subsequences between successive periods). At the first level, the shuffled text losses all kinds of correlation between consecutive occurrences of each symbol. The corresponding random walk is purely diffusive, and $\alpha = 1$. At the other two levels, on the other hand, some linguistic information persists after shuffling. At the level of words, both word frequencies and their internal structure, as strings of letters, is preserved. At the level of sentences, syntax rules still hold in the shuffled text, but the relation between sentences —and, thus, the semantic contents— is lost.

Remarkably, already for shuffling at the level of sentences, the symbolic sequence was shown to loose its correlations. The lower curve in Fig. 3.1 stands for the case of the German *Bible* shuffled at that level, with a slope corresponding to the uncorrelated random walk. Similar results were obtained for the other texts, shuffled both at sentence and word levels. This implies that the main contribution to the long-range organization of texts

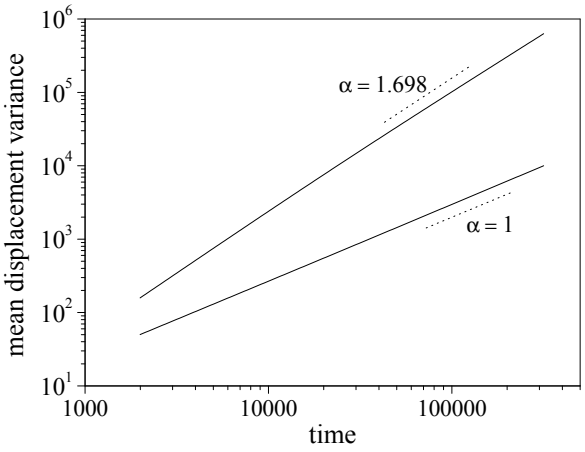

Figure 3.1: Upper curve: Mean displacement variance $\sigma^2(t)$ in the 32-dimensional random walk generated from a German edition of *The Bible*, as a function of time. Lower curve: Mean displacement variance for the German *Bible* shuffled at the level of sentences. Shuffling at the level of words produces essentially the same result. The dotted segments have the slopes indicated by their labels. Adapted from Ebeling and Neiman (1995).

—at least, to the degree discernable by the scaling of the random-walk displacement with the text length— comes from the correlation between sentences, i.e. by the specific ordering of grammatically consistent word sequences along the text. In other terms, the correlations detected by the random-walk representation of the text are mainly of semantic origin, and transcend the scope of syntax rules.

Note, from Fig. 3.1, that the anomalous power-law behavior of $\sigma^2(t)$ for *The Bible* is present from values of t of the order of a few thousands. The same is found for other long language samples. Although the power-law dependence of the mean displacement variance prevents defining a typical scale along the text, because of its intrinsic scale-free nature, this feature is an indication that long-range correlations are already well developed for such text lengths, which correspond to several hundred words.

The above analysis was complemented by the calculation of the Hölder exponents D_q, defined by the relations

$$\sum_{k=1}^{32} \langle |y_k(t) - \langle y_k(t) \rangle|^q \rangle \sim t^{D_q}, \tag{3.3}$$

for $q = 2, 3, \ldots$, which characterize deviations from normal diffusive behavior in higher order moments of the random walk displacements. For $q = 2$, D_q coincides with the diffusion exponent α. For normal diffusion, the relation $D_q = q/2$ holds. Calculation of D_q for the random walks generated

from texts yielded, instead,

$$D_q \approx \gamma q, \tag{3.4}$$

for $q \leq 6$, with $\gamma \approx 0.85$ for the German Bible, 0.59 for Grimm's *Tales*, and 0.62 for *Moby Dick*. To a good approximation, it is found that $\gamma = \alpha/2$, which means that higher order moments of the distribution of displacements exhibit the same scaling as the second moment σ^2.

The fact that long-range correlations in language samples already disappear upon shuffling the text at the level of sentences and words —which, as discussed above, is an indication that the overall organization of language is more related to the distribution and ordering of words than to the arrangement of letters— suggests that the representation of texts as letter sequences may not be the better choice from the viewpoint of quantitative linguistics (see also Secs. 4.1 and 4.2). The coding of words in a particular alphabet or phoneme set is, to a large extent, irrelevant to the linguistic structure of communication. A more adequate mapping of language onto a symbolic sequence should identify symbols with the basic elements intrinsic to the communication process (Montemurro and Pury, 2002). The next section reviews a statistical study of language mapped, as above, on a random walk, but taking individual words as its basic meaningful units.

3.2 Fractal Features in Word Sequences

Keeping in mind the aim of statistically identifying a stream of language with a random walk, Montemurro and Pury (2002) proposed to represent texts as symbolic sequences where each symbol was associated with a different word. In order to define the random walk, each different word in the text under study was first assigned a numeric value, namely, its rank r in Zipf's frequency list (see Sec. 2.1). The symbolic sequence was thus transformed into a discrete "time" series $r(t)$, corresponding to the rank of the word a place t in the text. Defining the average and the variance of $r(t)$ over the whole sequence as

$$\langle r \rangle = \frac{1}{T}\sum_{t=1}^{T} r(t), \quad \sigma_r^2 = \frac{1}{T}\sum_{t=1}^{T}(r(t) - \langle r \rangle)^2, \tag{3.5}$$

where T is the total text length measured in number of words, it is possible to introduce a rescaled time series,

$$\xi(t) = \frac{r(t) - \langle r \rangle}{\sigma_r}, \tag{3.6}$$

with zero mean and unitary mean square dispersion.

The rescaled series $\xi(t)$, which makes it possible to compare word sequences of different lengths and lexicon sizes, was used to define a one-dimensional random walk. At each time step t, the random walker jumps a

distance $\xi(t)$, so that its position at time t is

$$y(t) = y(0) + \sum_{u=1}^{t} \xi(u), \tag{3.7}$$

where $y(0)$ is the initial position. Montemurro and Pury (2002) analyzed the signal $y(t)$ generated from several language corpora, focusing on its statistical fractal properties. Specifically, they evaluated the Hurst exponent (Feder, 1988), which characterizes the scaling of long-range correlations related to the self-similar properties of the time signal. The same technique had already been used to analyze fractality in "DNA walks" representing nucleotide sequences in the genetic code (Borovik *et al.*, 1994).

The Hurst exponent H is introduced as follows. Define first the detrended signal

$$D(u, t, \tau) = y(t + u) - y(t) - \frac{u}{\tau}[y(t + \tau) - y(t)], \tag{3.8}$$

with $0 \leq u \leq \tau$, which measures the variation of $y(t)$ from t to $t + u$ relative to a linear interpolation between t and $t + \tau$. Then, take the cumulated range of $D(u, t, \tau)$,

$$R(t, \tau) = \max_{0 \leq u \leq \tau} D(u, t, \tau) - \min_{0 \leq u \leq \tau} D(u, t, \tau), \tag{3.9}$$

and the variance of the random walker's steps over the same interval,

$$\sigma^2(t, \tau) = \frac{1}{\tau} \sum_{u=t}^{t+\tau} \xi^2(u) - \left[\frac{1}{\tau} \sum_{u=t}^{t+\tau} \xi(u) \right]^2. \tag{3.10}$$

It turns out that, for a time signal exhibiting self-similarity, the ratio between the cumulated range and the step mean-square displacement averaged over all the possible starting steps t behaves as

$$\left\langle \frac{R(t, \tau)}{\sigma(t, \tau)} \right\rangle \sim \tau^H, \tag{3.11}$$

with $H > 1/2$. For a time signal corresponding to normal (uncorrelated) diffusion, on the other hand, it can be analytically proven that

$$\lim_{\tau \to \infty} \tau^{-1/2} \left\langle \frac{R(t, \tau)}{\sigma(t, \tau)} \right\rangle \sim \text{constant}, \tag{3.12}$$

which amounts to having $H = 1/2$. A Hurst exponent $H > 1/2$ corresponds to a random walk where long-range correlations favor persistent behavior, so that motion in a given direction promotes the future occurrence of displacements in the same direction.

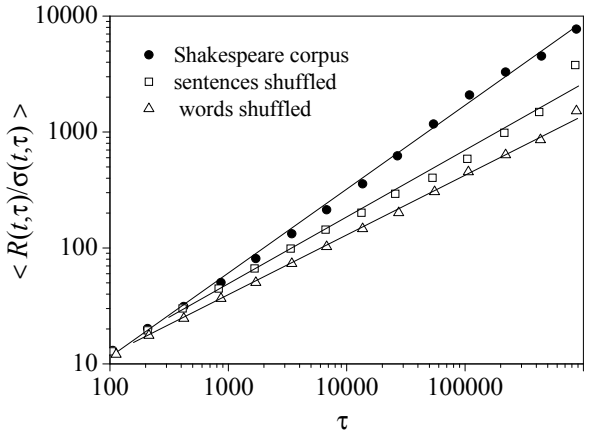

Figure 3.2: Average ratio of the cumulated range $R(t,\tau)$ and the mean square dispersion $\sigma(t,\tau)$ over intervals of length τ, for the Shakespeare corpus. Full dots stand for the results obtained for the original texts, while empty symbols correspond to shuffled texts at the level of sentences and words. Adapted from Montemurro and Pury (2002).

	H	sentences shuffled	words shuffled
Shakespeare	0.69	0.57	0.52
Dickens	0.74	0.57	0.52
Darwin	0.75	0.58	–

Table 3.1: First column: The Hurst exponent H of random walks generated from literary corpora by three authors, as described in the text. The second and third columns show H for a shuffling of the corpora at the level of sentences and words, respectively. Adapted from Montemurro and Pury (2002).

The quantity $\langle R(t,\tau)/\sigma(t,\tau)\rangle$ was computed as function of τ for the random-walk signals generated from three text corpora, each of them consisting of the concatenation of a number of works by a single author: 36 plays by William Shakespeare, 56 novels and other literary writings by Charles Dickens, and 11 books by Charles Darwin. Each corpus was several hundred thousands words in length, with a lexicon of a few tenth thousands different words. Full dots in Fig. 3.2 show the results for the Shakespeare corpus, in a log–log plot. The power-law behavior, here represented by linear profiles, is already well developed for τ of the order of a few hundreds. Similar results were obtained for the texts by Dickens and Darwin. The slope of a linear fitting of $\langle R(t,\tau)/\sigma(t,\tau)\rangle$ as a function of τ in this kind of plots provides a direct measure of the Hurst exponent H. The first numeric column in Table 3.1 shows the result of such fitting for the three corpora. In all cases, the values of H confirms the presence of long-range correlations.

The left panel of Fig. 3.3 shows an actual representation of position versus time for the random walk generated from the Shakespeare corpus. The long spans of average motion in either direction are indicative of the persistent behavior induced by long-range correlations. For comparison, the right panel shows two random walks generated from a stochastic sequence with $H = 0.5$ and 0.7, by the method of successive random additions (Voss, 1985). The latter exhibits the same persistent runs as the Shakespeare random walk.

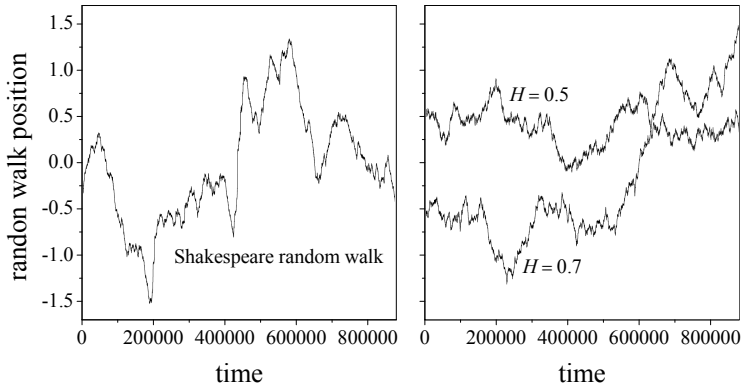

Figure 3.3: Left panel: Walker's position $y(t)$ as a function of time for the random walk generated from the Shakespeare corpus. Right panel: Two random walks generated from a stochastic sequence, with Hurst exponent $H = 0.5$ and 0.7. Adapted from Montemurro and Pury (2002).

As an interesting additional experiment, Montemurro and Pury (2002) considered, for each corpus, a subsequence obtained by deleting from the original texts all the words whose Zipf ranks were outside the interval $100 < r < 2000$. Namely, they excluded both the most frequent and the most rare words of each corpus. The latter, in particular, comprise thousands of different words. Remarkably, the Hurst exponents for these subsequences coincided, up to the empirical error, with those of the full corpora. The authors concluded that long-range correlations are, on the whole, induced by the distribution of neither the most frequent words nor the least used.

Finally, following the same procedure as Ebeling and Neiman (1995) had used for letter sequences (see Sec. 3.1), the word sequences corresponding to the three corpora were shuffled at two levels. At the first level, the internal structure of individual sentences was preserved, thus maintaining the grammatical order of words. At the second, word ordering was fully destroyed. Empty symbols in Fig. 3.2 stand for the average ratio $\langle R(t,\tau)/\sigma(t,\tau)\rangle$ as a function of τ for the shuffled Shakespeare corpus, and the two last columns in Table 4.1 show the Hurst exponents H of the resulting sequences for the three corpora. In contrast with Ebeling and Neiman's results, whose mea-

surements of diffusion coefficients were not able to discern between shuffling of letter sequences at the level of sentences and words, the Hurst exponents turned out to be different for shuffling of word sequences at each level. However, the significant decrease of H between the original sequences and those shuffled by preserving the ordering of words inside sentences, confirms the main conclusion that the rules of syntax are not enough to induce long-range correlations is language. A substantial portion of linguistic organization occurs beyond the scope of sentences, and must be ascribed to semantic structures associated with the topical development of the discourse.

3.3 Word Burstiness: A Key to Keywords

As we have seen in the preceding two sections, long-range correlations in the successive occurrences of the individual elements that constitute a stream of language, be they letters or words, provide strong evidence that some statistical regularities extend far beyond the scope of syntax rules. Linguistic structures emerge at the larger scales where the mutual interplay of the meaning of words gives rise to the semantic contents of writing or speech.

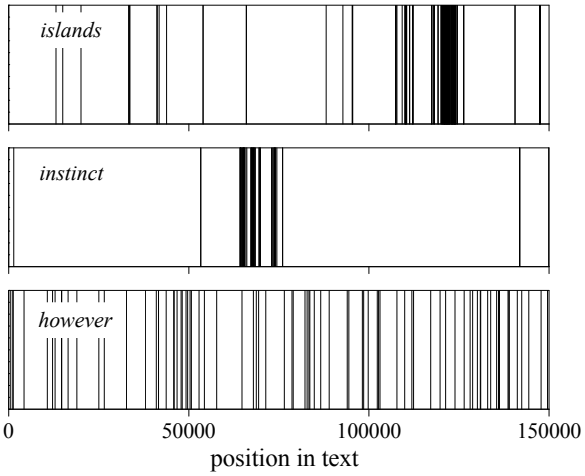

Figure 3.4: Position along Charles Darwin's *On the Origin of Species* of the words *islands* (143 occurrences), *instinct* (67 occurrences), and *however* (82 occurrences). The total length of the text is slightly above 150000 words. See also Herrera and Pury (2008) and Montemurro and Zanette (2009).

Still, careful comparison of the distribution of different words along a text reveals quite disparate situations. Consider, for instance, Charles Darwin's *On the Origin of Species* —the treatise where the great naturalist synthesized a life of observation and thinking about the evolution of the living world. In Darwin's book, the word *islands* occurs 143 times. Not unexpectedly, most of these occurrences come in the two chapters where

Darwin expounds how new species may develop by the isolation of animals and plants in different environments. The top panel of Fig. 3.4 represents, as vertical bars, the successive positions of *islands* along the book. The clusters around position 120000 are situated inside those two chapters. Similarly, the middle panel shows the positions of the word *instinct*. Of its 67 occurrences, more than 50 belong to the chapter titled, precisely, *Instinct* (Montemurro and Zanette, 2009). On the other hand, the 82 occurrences of the word *however*, plotted in the lower panel, display a much more homogeneous distribution. This is, again, not unexpected, in view that the role of *however* in English —as a connector between phrases which express contrasting ideas— is mainly functional, and does not bear any specific relation to the subjects of the book or of any of its parts.

It turns out that heterogeneity in the distribution of word occurrences —with accumulation, or clustering, in some parts of the text and more sparse appearances in other parts— is a general feature of those words which are directly related to the topics developed in a book or, more generally, in any long, meaningful piece of written or spoken discourse. The accumulation of the occurrences of topical words in certain parts of the discourse, which is very graphically referred to as *burstiness*, seems therefore to be a generic property of the way humans organize word sequences to convey complex information. In addition, this property provides an invaluable tool to detect highly meaningful words, i.e. keywords, just on the basis of their distribution along a text —a task of much importance in automatized analysis of language and information retrieval (Indurkhya and Damerau, 2010).

A convenient way to give a quantitative characterization of burstiness in word distribution is to study the statistics of the number of words between two consecutive occurrences of each word w_i. Let us call this number the *distance* between the two occurrences of w_i, and denote it by x_i. If the total number of occurrences of w_i is n_i, and the probability of finding w_i at a given place in the text were the same all over the text —namely, if there were no correlations between successive occurrences of the word in question— the probability of the distance x_i would be given by the geometric distribution $p(x_i) = f_i(1 - f_i)^{x_i}$ $(x_i = 0, 1, 2, \dots)$, with $f_i = n_i/T$. This probability describes a random, but uniform, distribution of occurrences of w_i along the text, with an average distance between successive occurrences given by the inverse of the word's frequency: $\langle x_i \rangle = f_i^{-1}$. If f_i is not too large, the geometric distribution is well approximated by the Poissonian form

$$p(u_i) = \exp(-u_i), \tag{3.13}$$

where $u_i = x_i/\langle x_i \rangle$ is a rescaled variable, given by the ratio between the distance x_i and its mean value. Working with this rescaled variable has the advantage that the form of $p(u_i)$ for a uniform random distribution of word occurrences, Eq. (3.13), is the same for all words, independently of their individual frequencies f_i.

For words in a real text, $p(u_i)$ can be evaluated empirically by measuring the distances between consecutive occurrences. Differences with the Poissonian form are an indication of departure from the random uniform distribution. Analyzing several texts in English, which comprised novels, theatrical plays, and scientific and philosophical treatises, Álvarez Lacalle *et al.* (2006) showed that for topical words, closely related to the main subjects developed in each book, the autocorrelation between word positions along the text decreases as a power law. This is consistent with a similar decay in the distribution of distances, $p(u_i) \sim u_i^{-\gamma_i}$, which differs from the Poissonian profile.

If, for small distances, $p(u_i)$ is larger than the Poissonian distribution, the occurrences of the word in question tend to cluster in some regions and to become depleted in other zones. As some authors put it, in this case the word exhibits "self-attraction" (Ortuño *et al.*, 2002). This is a direct clue to burstiness. In the opposite case, $p(u_i)$ is relatively depleted for small u_i, and the word "repels" itself. Since, however, the number of occurrences is fixed and the text has finite length, the distance between occurrences cannot grow indefinitely. Rather, the consecutive appearances of the word tend to be more equally spaced than for the Poissonian distribution. In the extreme (and unlikely) case where all the occurrences were equally spaced, the distribution of distances would be infinitely concentrated around the average value: $p(u_i) = \delta(u_i - 1)$.

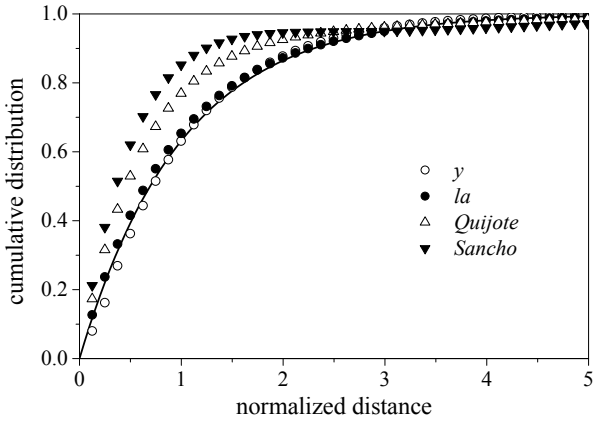

Figure 3.5: Schematic representation of the cumulative distribution of distances between consecutive appearances of four words from Miguel de Cervantes's *Don Quijote*: *y* ("and"), *la* (singular feminine form of "the"), *Quijote*, and *Sancho*. For each word, distances are normalized by their mean value over the text. The curve represents the cumulative Poisson distribution $P(u)$ of Eq. (3.14). Adapted from Ortuño *et al.* (2002).

Figure 3.5 shows, as a curve, the cumulative distribution

$$P(u) = \int_0^u p(u)du' = 1 - \exp(-u) \tag{3.14}$$

for uniformly distributed words, and the empirical evaluation of the same distribution for four words of Miguel de Cervantes's *Don Quijote*. In agreement with the above discussion, the functional words *y* ("and") and *la* (singular feminine definite article, "the") follow closely the cumulative Poisson distribution. *Quijote* and *Sancho*, on the other hand, are the proper names of the novel's two main characters, and are therefore highly topical to the book. Accordingly, the cumulative distributions of the distance between successive occurrences differ sensibly from the Poissonian form, disclosing a more heterogeneous arrangement of those two words along the text. The relatively large values of the empirical cumulative distributions for small distances reveal the tendency of their occurrences to aggregate into clusters.

Early proposals to quantitatively characterize burstiness were based on the description of the empirical distributions of distances between word occurrences by means of "Poisson mixtures" (Church and Gale, 1995; Katz, 1996). A simpler and more compact characterization was suggested by Ortuño *et al.* (2002), who introduced the standard deviation of u_i,

$$\sigma_i = \langle (u_i - \langle u_i \rangle)^2 \rangle^{1/2} \tag{3.15}$$

as a burstiness index for word w_i. While the Poissonian distribution has $\sigma_i = 1$, "self-attracting" words show less homogeneous arrangements, and thus $\sigma_i > 1$. If, on the other hand, a word is more evenly spaced than for the Poissonian distribution, its burstiness index is $\sigma_i < 1$.

	σ_i	n_i
Jesus	24.18	983
Christ	18.42	571
Paul	11.56	162
disciples	10.88	244
Peter	10.17	164
Joab	10.03	145
faith	9.34	247
Saul	9.17	420
Absalom	9.12	108
John	9.03	137

Table 3.2: The ten words with the largest burstiness index σ_i in *The Bible* (in English). For each word, the number of occurrences n_i is also given. Adapted from Ortuño *et al.* (2002).

Ortuño *et al.* (2002) computed the burstiness index for the words of several texts, including an English version of *The Bible*. Table 3.2 shows

the ten words with the largest values of σ_i in this book, along with their number of appearances. Note first that the values of σ_i are sensibly above one. All these words, thus, exhibit a high degree of burstiness. Moreover, there is no obvious relation between the burstiness index and the number of appearances. Frequent words can display low burstiness, and *vice versa*. The most striking fact in the words of Table 3.2 is however their high topicality with respect to *The Bible*. Most of them are the proper names of some of its main protagonists. The only two common names in the list, *disciples* and *faith*, are also highly relevant to the book's subjects.

The same study included a comparison of the values of σ_i for all the words in a given book, with those of the words in the book's glossary — a collection of terms purposely chosen to serve as a reference to the main topics covered by the work. It was found that the burstiness of the words in the glossary was, on the average, larger than twice the mean burstiness all over the book's lexicon. These empirical results led the authors to propose using the burstiness index as a computationally cheap method to extract keywords from a text in electronic format. The selection of words on the basis of σ_i should be complemented by fixing a lower bound in the number of occurrences: very unfrequent words can also have relatively high burstiness indices, but could be specific to irrelevant parts of the text. Zhou and Slater (2003), moreover, pointed out the need to take into account not only the distance between consecutive appearances, but also the distance between the first and the last appearances from the beginning and the end of the text, respectively. In fact, a word with $\sigma_i \approx 1$, but confined to a small part of the text, may still be a very topical one.

Also with the aim of introducing a tool for automatic keyword detection, Herrera and Pury (2008) proposed an alternative characterization of word burstiness in terms of an entropy-like measure. The method was based on studying the heterogeneity of the distribution of individual words over a prescribed partition of the text in question (Montemurro and Zanette, 2002a). First, they divided the text of length T into M parts, of lengths $T_1, T_2, \ldots, T_M$. The frequency of each word w_i in part m was written as $f_i^m = n_i^m/T_m$, where n_i^m was the number of occurrences of w_i in m. The quantities

$$p_i^m = \frac{f_i^m}{\sum_{k=1}^{M} f_i^k} \tag{3.16}$$

define a normalized distribution over the M parts of the text: for each word, $\sum_m p_i^m = 1$. The entropy associated with this distribution is[1]

$$h_i = -\sum_{m=1}^{M} p_i^m \log_M p_i^m. \tag{3.17}$$

[1]Here, in contrast to Sec. 1.3, we use for the entropy the lower-case notation h_i to emphasize that it is a quantity defined for each single word, and not for the set of all the words in the text. A similar notation will later be used in Chap. 4.

The extreme values of this entropy, $h_i = 0$ and 1, are attained when w_i appears in only one part, and when its number of occurrences in each part m is proportional to T_m, respectively.

It can be shown that, when the parts are equal in length, the expected value of the entropy h_i for a word whose n_i occurrences are distributed uniformly at random along the text is $h_i = 1-(M-1)/2n_i \ln M$ for large n_i (Montemurro and Zanette, 2002a). On the basis of this result, Herrera and Pury (2008) suggested to characterize the heterogeneity in the distribution of word w_i over the text partition by means of the index

$$E_i = \frac{2 \ln M}{M-1} n_i (1 - h_i). \tag{3.18}$$

For equal parts and a random uniform distribution, $E_i \approx 1$. Words with heterogeneous distributions, on the other hand, are expected to have larger values of E_i.

The index E_i was calculated for all the words in Darwin's *On the Origin of Species*. The partition of the text was chosen to coincide with its sixteen chapters. In order of decreasing E_i, the twenty words with the largest indices were the following: *hybrids*, *I*, *sterility*, *islands*, *species*, *forms*, *varieties*, *instincts*, *breeds*, *fertility*, *formations*, *crossed*, *selection*, *organs*, *characters*, *nest*, *instinct*, *rudimentary*, *formation*, and *genera*. Except for *I*, which appears second in the list, it is clear that all these words are highly topical to the book.

These results reinforce the notion that heterogeneity in the distribution of a word is a signature of topicality and, at the same time, show that entropy-like measures are also suitable to describe such heterogeneity. While the introduction of a partition adds in principle a largely arbitrary set of parameters (the part lengths T_m), thus weakening the generality of quantitative results, applying the method to a natural partition of the text —for instance, its division into chapters— may enhance the significance of the heterogeneity in the distribution of certain words with respect to that partition. Frequent words specific to a given part, in particular, would be assigned maximal values of the index E_i.

On the other hand, the presence of the non-topical word *I* high in the above list puts forward a caveat on the use of this kind of methods for keyword detection. The large value of E_i for the word *I* is related to the switching of Darwin's style along *On the Origin of Species*, between scholarly discourse and first-person narrative (Herrera and Pury, 2008), which naturally makes the distribution of *I* very heterogeneous. However, methods such as those described above —which are purely based on a statistical characterization of burstiness— cannot discern between different sources of heterogeneity. As other algorithms for automatic text analysis, thus, they should be regarded as potential components of a battery of tools whose efficiency depends, precisely, on being used in combination with each other.

As a closing note, it is interesting to mention that the burstiness measure defined in Eq. (3.15) has also been applied to DNA sequences (Ortuño *et al.*, 2002). It is well known that deoxyribonucleic acid (DNA) is a biological macromolecule that stores the genetic information on which the ontogenic development of all living organisms is based. Structurally, it consists of a sequence of nucleotides of four types (A, T, C, and G), whose specific ordering encodes the information conveyed by the sequence. As an information carrier, the four-symbol DNA code is appealingly reminiscent of language. It should be recalled, however, that not all of the DNA sequence bears genetic information: large non-coding sectors, whose function is not yet understood, are intercalated between coding sections (Klug and Cummings, 1997).

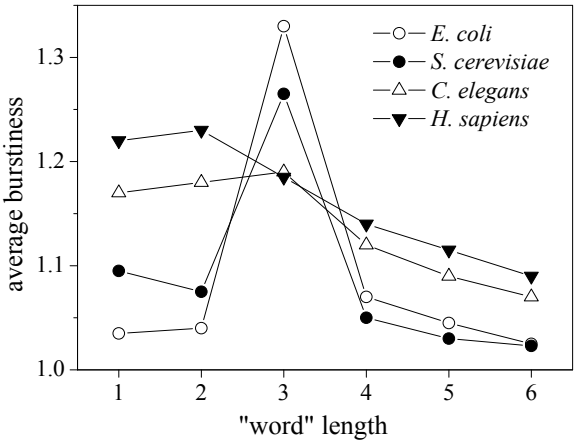

Figure 3.6: Average burstiness index σ for DNA "words" in four biological species, as a function of the "word" length l. A "word" is defined as an array of l consecutive nucleotides in the DNA sequence. Adapted from Ortuño *et al.* (2002).

Ortuño *et al.* (2002) divided long samples of DNA sequences of several biological species into non-overlapping "words" consisting of l contiguous nucleotides. There are 4^l such "words." For each $l = 1, 2, \ldots, 6$ —and starting at l consecutive places in the sequence, to take into account all its possible partitions— the burstiness index σ_i was calculated for each "word" and then averaged over the whole "lexicon." Figure 3.6 displays the average burstiness σ as a function of l for four living species: the bacterium *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, and the hominid *Homo sapiens*. In all cases, and for all values of l, the index is larger than one, revealing the ubiquity of burstiness in DNA. The values of σ, however, are well below those found for some topical words in language samples (see Table 3.2).

The most interesting feature is, in any case, the peak in the burstiness

index at $l = 3$ (except for *H. sapiens*). Nucleotide triads have been identified as the most elementary units of genetic information —the *codons*— each of them controlling the process of production of a different aminoacid (Klug and Cummings, 1997). The coding sections of DNA are arrays of contiguous codons. It is noticeable that the position of the maximum of σ coincides with the length of these basic "words" of the genetic code. The fact that the peak becomes less significant as the complexity of the organisms grows has been associated with their increasing relative contents of non-coding DNA, which reaches more than 98% for *H. sapiens*. The results shown in Fig. 3.6 for our species suggest that, if non-coding DNA stores any information, one- and two-nucleotide "words" might be playing a relevant role in this function.

Chapter 4

Order, Entropy, and Information in Written Texts

From the perspective of quantitative linguistics, measuring the amount of information conveyed by human language is a fascinating challenge. This quest —which is directly related to the quantification of redundancy in speech and writing, of the constraints imposed by the function of language as a means of communication, and of the degree of order in linguistic patterns— is best dealt with using the tools of information theory. It is in this conceptual framework, in fact, that a generic communication channel can be mathematically characterized, gauging the roles of source, transmitter, receiver, and the intervening noise. Information theory was applied to the analysis of language from its very inception by Claude Shannon, who presented the first estimation of the entropy of written English —as an inverse measure of its degree of redundancy, predictability, and organization— in 1951. This classical result immediately elicits many intriguing questions. Are all human tongues, from a statistical viewpoint, equally organized? Does the information contents of a text depend on its literary genre or style? Was the emergence of redundancy favored or hindered along language evolution? Most of these questions remain, to a large extent, unexplored. In the last few decades, however, a renewed interest in the applications of information theory to human language —also inspired by the affinity with other natural "languages," such as the genetic code stored in DNA (see Sec. 3.3)— has led to novel results on the statistical properties of grammatical and semantic patterns.

We begin this chapter with a review of the classical definition and calculation of the entropy of language, as conceived by Shannon, which regards language samples as partially predictable sequences of letters. Then, we focus on information-theoretical approaches which, more significantly from a linguistic viewpoint, analyze the statistical properties of word sequences. In this framework, it has been shown that the amount of information which

is lost when shuffling the words in a sample of language is the same in very distant tongues, thus constituting a kind of quantitative linguistic universal. In turn, the mutual information between the bursting distribution of topical words and a partition of the text to which they belong, makes it possible to define a typical scale —of the order of several hundred to a few thousand words— associated with the unfolding of semantic contents. Finally, not only topicality but also the grammatical function of words is shown to influence the uniformity of their usage across literary texts.

4.1 Shannon's Evaluation of the Entropy of Printed English

"Can we define a quantity which will measure, in some sense, how much information is 'produced' by... a [stochastic] process, or better, at what rate information is produced? Suppose we have a set of possible events whose probabilities... are known but that is all we know concerning which event will occur. Can we find a measure of how much 'choice' is involved in the selection of the event or of how uncertain we are of the outcome?" With these questions, Claude Shannon motivated, in his foundational papers on information theory, the introduction of the entropy of a source of information (Shannon, 1948a,b). At that stage, he conceived an information source as the emitter of a message in the form of an ordered sequence of discrete symbols, able in principle to produce infinitely long sequences.

According to Shannon, the entropy per symbol —or *entropy rate* (Cover and Thomas, 2006)— of the information source is given by the limit

$$h = \lim_{n\to\infty} \frac{H_n}{n}, \tag{4.1}$$

with

$$H_n = -\sum_{\{s_1,s_2,\ldots,s_n\}} p(\{s_1, s_2, \ldots, s_n\}) \log_2 p(\{s_1, s_2, \ldots, s_n\}). \tag{4.2}$$

Here, $p(\{s_1, s_2, \ldots, s_n\})$ is the probability of occurrence, along the symbol sequence, of the n-symbol subsequence (or *block*) $\{s_1, s_2, \ldots, s_n\}$. The sum runs over all the possible subsequences of length n; if the number of different symbols is V, there are V^n such subsequences. In Eq. (4.2), H_n is the entropy associated with the probability distribution $p(\{s_1, s_2, \ldots, s_n\})$ over the set of V^n subsequences. The choice of the logarithm base amounts to fixing the units in which the block entropies H_n and the entropy per symbol h are measured. Using $\log_2$ defines such units as bits and bits per symbol, respectively (see Sec. 1.3). When the sequence is a written text where each letter is identified with a symbol, in Shannon's words, if the text *"is translated into binary digits (0 or 1) in the most efficient way, h is the*

number of binary digits (i.e. bits) *required per letter of the original language"* (Shannon, 1951).

Since h integrates knowledge about the probabilities of all possible symbol arrays of any length, it bears information on the statistical correlations of the sequence —and, thus, on its organizational structure— at all scales. Large values of the entropy are obtained when the probabilities of different subsequences of a given length are similar, and therefore correspond to high degrees of uncertainty. In this situation, the signal redundancy is small, and the observation of any given subsequence provides little information about the rest of the signal. Small entropies, on the other hand, are associated with repetitive, redundant, highly predictable sequences. Shannon gave a quantitative definition of redundancy as

$$r = 1 - \frac{h}{h_{\max}}, \tag{4.3}$$

where $h_{\max} = \log_2 V$ is the maximum entropy per symbol attainable by a sequence built up from the collection of V symbols. This corresponds to a random sequence where all symbols occur with the same probability, uncorrelated to each other. The redundancy r is a measure of the fraction of the sequence that is determined by its structural constrains or, as Shannon put it, the part that cannot be *"chosen freely."*

In an alternative formulation of the same definition for the entropy, Shannon introduced the quantities

$$F_{n+1} = -\sum_{\{s_1,\ldots,s_n\}} \sum_{\{s_{n+1}\}} p(\{s_1,\ldots,s_n,s_{n+1}\}) \log_2 p(s_{n+1}|\{s_1,\ldots,s_n\}), \tag{4.4}$$

where $p(s_{n+1}|\{s_1,\ldots,s_n\})$ is the conditional probability of occurrence of symbol s_{n+1} given that it was preceded by the subsequence $\{s_1,\ldots,s_n\}$. It can be proven that $F_{n+1} = H_{n+1} - H_n$ and, from here, that

$$h = \lim_{n\to\infty} F_n. \tag{4.5}$$

The quantity F_{n+1} measures the degree of uncertainty in the $(n+1)$-th symbol, under the assumption that one knows the precedent n symbols. Therefore, it provides an inverse measure of the predictability of the sequence, given a certain amount of information about its foregoing constitution.

While Shannon's work was aimed at giving a mathematical basis to the description of generic communication processes —including, in particular, those occurring between or mediated by artificial (electrical) devices— his references to human language were unavoidable (Shannon, 1948a). Three years after his first papers on the subject, he published his own empirical results on the entropy of written (or *"printed"*) English (Shannon, 1951). In this study, English texts were viewed as sequences of 27 symbols, including

the 26 letters of standard English and the space between words. Any other written symbols were disregarded.

Shannon estimated the entropy on written English on the basis of its predictability, namely, from the evaluation of the quantities F_n defined from Eq. (4.4). For the first few values of n, he relied on preexisting tabulations of the frequencies of single letters, digrams (two-letter subsequences) and trigrams (three-letter subsequences) in large corpora of English texts, which yielded $F_1 = 4.03$, $F_2 = 3.32$, and $F_3 = 3.1$ bits per letter. For longer subsequences, however, tabulations were not available, and it was necessary to resort to different means to evaluate predictability beyond three letters. Shannon's ingenious solution to this problem was to devise a method that exploited the fact that *"anyone speaking a language possesses, implicitly, and enormous knowledge of the statistics of language. Familiarity with the words, idioms, clichés and grammar enables him to fill in missing or incorrect letters in proof-reading, or to complete an unfinished phrase in conversation"* (Shannon, 1951).

With this idea in mind, Shannon designed two experiments, involving human subjects.[1] In the first one, a text unfamiliar to the subject was selected, and the subject was asked to guess the first letter in the passage. If the guess was correct, he or she was so informed and proceeded to guess the following letter (or space). If it was not, the subject was told the correct letter and then proceeded to the next guess. This procedure was continued through the whole text. The correct text up to each point was available to the subject for use in predicting the following letters. A typical result of this experiment, exemplified by a short sentence, was as follows:

T H E R O O M W A S N O T V E R Y L I G H T
- - - - R O O - - - - - - N O T - V - - - - - I - - -

The first line reproduces the original text. In the second line, dashes and letters indicate correct and incorrect guesses, respectively.

In this first experiment, out of a total of 129 letters, 89 (69%) were guessed correctly. Errors, not unexpectedly, occurred more frequently at the beginning of words and syllables. Shannon pointed out that, at first sight, it might seem that the second line contains much less information than the first. Actually, however, their information contents is the same, in the sense that the subject was able to build up the original text just with the information given in the second line. A subject's "mathematically identical" twin, who would respond exactly in the same way when faced to the same problem, should be able to recover the first line from the second. The redundancy of 69% ($r = 0.69$) found in this experiment was sensibly higher than the value obtained for written English in work previous to Shannon's,

[1] In view of Shannon's acknowledgements in his 1951 paper, the subjects were his wife, Mrs. Mary E. Shannon, and his collaborator, Dr. B. M. Oliver.

of about 50% (Shannon, 1948a).

The second experiment provided a more direct evaluation of predictability. As in the first one, the subject knew the text up to the current point and was asked to guess the next letter. If the guess was wrong, he or she was told so and was asked to guess again. This was repeated until the correct letter was found, and the number of guesses needed to find the correct letter was recorded at each step. A typical result was as follows:

T	H	E	R	E		I	S		N	O		R	E	V	E	R	S	E		O	N		A				
1	1	1	5	1	1	2	1	1	2	1	1	15	1	17	1	1	1	2	1	3	2	1	2	2			
M	O	T	O	R	C	Y	C	L	E		A		F	R	I	E	N	D		O	F		M	I	N	E	
7	1	1	1	1	4	1	1	1	1	1	3	1	8	6	1	3	1	1	1	1	1	1	1	1	1	1	1
F	O	U	N	D		T	H	I	S		O	U	T		R	A	T	H	E	R							
6	2	1	1	1	1	1	1	2	1	1	1	1	1	1	4	1	1	1	1	1	1						
D	R	A	M	A	T	I	C	A	L	L	Y		T	H	E		O	T	H	E	R		D	A	Y		
11	5	1	1	1	1	1	1	1	1	1	1	1	6	1	1	1	1	1	1	1	1	1	1	1	1		

Here, the second line indicates the corresponding number of guesses for each letter (or space).

To make the experiment's results statistically more significant, each subject was required to guess the text, letter by letter, of one hundred samples of literary English —taken at random from D. Malone's *Jefferson the Virginian*— fifteen letters in length each. Additionally, a similar test was performed in which 100 letters were known to the subject and the 101-st letter was to be guessed. The fraction of times q_k^n that the n-th letter was correctly guessed at the k-th guess was recorded.

Shannon noticed that the quantities q_k^n are linked to the conditional probabilities $p(s_{n+1}|\{s_1,\ldots,s_n\})$ of Eq. (4.4) by the relation

$$q_k^{n+1} = -\sum_{\{s_1,\ldots,s_n\}} p(s_k|\{s_1,\ldots,s_n\}), \tag{4.6}$$

where the sum runs over all the subsequences of length n and s_k is the k-th most likely letter to follow each particular subsequence. This relation makes it possible, through rather involved algebraic arguments, to find a connection between the quantities q_k^n and F_n. Specifically, it is possible to show that

$$\sum_{k=1}^{27} k(q_k^n - q_{k+1}^n)\log_2 k \leq F_n \leq -\sum_{k=1}^{27} q_k^n \log_2 q_k^n, \tag{4.7}$$

which fixes lower and upper bounds for F_n.

Figure 4.1 shows a plot of the bounds for F_n as functions of n, computed from the results of Shannon's second experiment. Perhaps the most significant feature emerging from these results is that, while F_n seems to have attained a rather stable value for $n = 15$, there is still a substantial difference with the estimation of F_{100}. Throughout the range not covered by the experiment, F_n decays from around 2 bits/letter to about half that

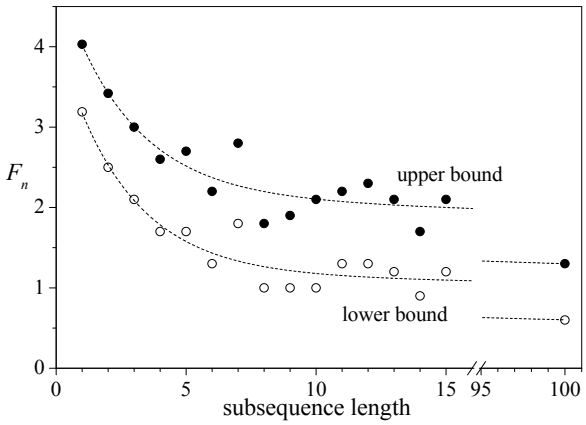

Figure 4.1: Lower and upper bounds for the quantities F_n measured in bits per letter, as functions of the subsequence length n, estimated for written English from inequalities (4.7). The limit of F_n for $n \to \infty$ gives the entropy per symbol of a sequence. Dashed lines are plotted as a guide to the eye. Adapted from Shannon (1951).

value. This reveals the existence of nontrivial structures in the letter sequence, extending for at least several tenths of symbols. The estimation of the entropy of written English arising from the experiment is of about 1 bit per letter. The corresponding redundancy is just below 80%.

Since Shannon's times, several different procedures have been used to obtain other estimates of the entropy of English. Cover and King (1978) proposed a gambling game to evaluate predictability in language, again involving human subjects, and found $h \approx 1.3$ bits/letter. This method however, was criticized on the basis of the tendency of humans to overestimate the probability of very unlikely events (Schwartz and Reisberg, 1991) —a critique that may have also been applied to Shannon's experiments. Non-human methods included text compression based on prediction from partial matching (PPM), which was applied to English texts of several genres and styles, yielding $h \approx 1.5$ bits/letter (Teahan and Cleary, 1996). Kontoyiannis *et al.* (1998), in turn, used elaborate nonparametric entropy estimators and got, for four novels by Jane Austin, $h = 1.8$ bits/letter. Other authors (Ebeling and Poschel, 1994) complemented the calculation of the entropy with a detailed analysis of scaling properties in the convergence of the normalized block entropies H_n/n towards h, Eq. (4.1), as n tends to infinity.

While all these studies gave values for the entropy per letter in written English within a well defined range of 1 to 2 bits/letter, it cannot be said that they yielded a quantitatively very consistent, accurate result. The discrepancies can in part be ascribed to methodological differences but, surely, the fact that they were applied to different kinds of texts also played a role.

Probably none of the English corpora involved in these studies was, from a broad statistical viewpoint, really representative of their language.

There is however another major drawback in the above approaches to the definition of the entropy of written language, which jeopardizes their interest for linguistics. Considering a text as a sequence of letters (and other characters) brings into the calculation of entropy a contribution from the specific order in which letters are combined to form words, which has essentially no relevance to grammar (see also Sec. 3.1). Syntactic structures, as well as longer-ranged linguistic patterns, are expected to be largely impervious to the organization of letters inside words. Lexical similarities between distant languages, with quite disparate grammatical rules, suggest that the two aspects have had independent evolutionary origins. In the following sections, we review several information-theoretical studies of the statistics of language patterns which consider written texts as sequences of words, instead of letters. This linguistically more compelling point of view should help to better discern between the different contributions to the information conveyed by language.

4.2 The Information Stored in Word Ordering

In the framework of information theory, the convenience of viewing written texts as word sequences —instead of character sequences, as in the preceding section— is well illustrated by the following simple experiment. Suppose that, in a given text, each different word is substituted by a number, for instance, its rank in the Zipf frequency list (see Secs. 2.1 and 3.2). Obviously, this "translation" will leave the message communicated by the text unaltered. It suffices to have the proper "dictionary" to fully recover the original text from the numeric sequence. On the other hand, the entropy per symbol of this numeric sequence is expected to substantially differ form that of the original sequence of letters. Indeed, the amount of information associated with the specific ordering of letters inside words will not contribute to the former.

This observation recalls that language structures occur at many levels of organization, from the internal morphology of individual words to the long word sequences that define the contextual elements which endow the message with intelligibility and meaning. While all these levels contribute to the degree of order of language, their linguistic roles are not the same. The inflection of words —namely, the different word forms that derive from a single root through conjugation or declension— defines, at the shortest-range level, grammatical categories such as person, number, gender, and tense. At an intermediate level, the rules of syntax determine the relative order of words inside a phrase or sentence. In some human tongues, the mutual position of words is crucial to discern between subjects, actions, and

objects, while others compensate a larger flexibility in syntactic order with more complex inflection (Greenberg, 1963). Finally, semantic structures — those associated with the global meaning of a message— emerge at even larger ranges, involving several sentences or paragraphs, as the information contained in the message is unraveled and developed.

The fact that the words in the lexicon of a given language sample are not all used with the same frequency, as described by Zipf's law (see Chap. 2), reveals already a certain degree of nontrivial linguistic organization, amenable to quantification by means of an entropy-like measure. If, in a text of length T with a lexicon of V different words, the word w_i ($i = 1, 2, \ldots, V$) occurs n_i times, the quantity

$$H_Z = \log_2 \frac{T!}{n_1! n_2! \cdots n_V!} \tag{4.8}$$

is the entropy (measured in bits, see Sec. 1.3) associated with Zipf's distribution for the text in question. It equals the entropy of the ensemble of all the possible orderings of the text's words and can therefore be identified with the entropy of a random shuffling of the text. When the numbers n_i are generally large, by virtue of Stirling's formula (Abramowitz and Stegun, 1972), H_Z can be approximated as

$$H_Z \approx -\sum_{i=1}^{V} n_i \log_2 \frac{n_i}{T} = -T \sum_{i=1}^{V} f_i \log_2 f_i, \tag{4.9}$$

where $f_i = n_i/T$ is the frequency of appearance of w_i. Taking into account that, in a typical text or speech, there are many words which occur only a small number of times, the quality of the Stirling approximation must nevertheless be carefully assessed in each particular case. However, it turns out to be very convenient for analytical treatment. The quantity H_Z can be interpreted as the entropy of an infinitely long sequence of words —or of an infinitely large ensemble of finite sequences— where each word w_i appears with frequency (or probability) f_i.

The Zipf entropy H_Z is to be compared with $H_0 = T \log_2 V$, which stands for the entropy of a uniform probability distribution over the whole lexicon: $f_i = V^{-1}$ for all words w_i. The two quantities would coincide if all the words of the lexicon appeared in the text exactly the same number of times. The difference

$$H_0 - H_Z = T \log_2 V + T \sum_{i=1}^{V} f_i \log_2 f_i \tag{4.10}$$

is always positive, and is an estimation of the information associated with the actual distribution of word occurrences with respect to a "text" where all words are used with the same frequency. Note that the entropies H_Z and

H_0 are proportional to the text length T. It makes therefore sense to define such quantities as $h_Z = H_Z/T$ and $h_0 = H_0/T$, which provide measures of the entropy (or of the information) per word.

	h_Z	h_0	$h_0 - h_Z$
Æneid	12.27	14.02	1.75
Don Quijote	9.60	14.50	4.90
D. Copperfield	9.28	13.78	4.50

Table 4.1: Entropies per word, h_Z and h_0, as defined in the text, for *Æneid*, *Don Quijote*, and *David Copperfield*, and their difference $h_0 - h_Z$. The three quantities are measured in bits per word.

Table 4.1 shows the entropies per word, h_Z and h_0, and their difference, for Virgil's *Æneid*, Miguel de Cervantes's *Don Quijote*, and Charles Dickens's *David Copperfield* (see also Chap. 2). Owing to the comparable lexicon size of the three works, the respective values of h_0 are similar. On the other hand, h_Z is sensibly higher for *Æneid*. This is directly related to the fact that its Zipf distribution is much flatter than for the other two texts, as discussed in Chap. 2 in connection with Fig. 2.1. Accordingly, the information per word associated with the distribution of word occurrences, $h_0 - h_Z$, is smaller for the Latin poem.

At a higher organizational level, the structure of a language sample is determined by the particular manner in which words are arranged along their ordered sequence. The increment from the information contained in the mere collection of different words, each one considered with its respective number of occurrences, to the information contained in the specific word sequence that constitutes the actual text or speech, is a quantification of the cost of constructing a message useful to communication out of a disordered assemblage of words.

In order to capture the whole contribution of the correlations implicit in the sequence of words of a real text, according to Eqs. (4.1) and (4.2), its entropy should be obtained from the block entropies

$$H_n = - \sum_{\{w_1, w_2, \ldots, w_n\}} f(\{w_1, w_2, \ldots, w_n\}) \log_2 f(\{w_1, w_2, \ldots, w_n\}) \quad (4.11)$$

where, for each n, the sum runs over all the possible arrays of n words, $\{w_1, w_2, \ldots, w_n\}$. The quantity $f(\{w_1, w_2, \ldots, w_n\})$ is the frequency (i.e., an estimation of the probability) of appearance of each array over the whole text under study. From Eq. (4.1), the entropy per word to be assigned to the text is $h = \lim_{n\to\infty} H_n/n$. Any real text, however, is of course finite in length, so that the limit for $n \to \infty$ cannot be achieved. Moreover, the long-range correlations present in language (Ebeling and Poschel, 1994; Ebeling and Neiman, 1995) make a direct evaluation of the block entropies

H_n highly inaccurate. More precise methods, which deal efficiently with correlations, are based on the fact that the entropy of a symbolic sequence provides a lower bound for the length of a lossless compressed version of the same sequence (Cover and Thomas, 2006). In other words, the entropy can in principle be evaluated by finding the length of the sequence after having been compressed by means of an optimal algorithm. On the basis of the classical Lempel–Ziv compression algorithm (Ziv and Lempel, 1977), Kontoyiannis and coworkers, among others, have developed efficient entropy estimators, with robust convergence even in the presence of strong correlations (Kontoyiannis *et al.*, 1998; Gao *et al.*, 2008). These estimators advantageously replace the calculation of h by means of block entropies.

In an extensive study, applied to a large corpus consisting of 7077 texts,[2] the entropy per word h of each text was calculated using one of Kontoyiannis's estimators (Montemurro and Zanette, 2011). The texts were written in eight languages belonging to five different linguistic families and including a language isolate: English, French, and German, from the Indo-European family; Finnish, from the Finno-Ugric family; Tagalog, from the Austronesian family; Sumerian, which is unrelated to other known languages; Old Egyptian, from the Afro-Asiatic family; and Chinese, from the Sino-Tibetan family (Lewis, 2009). The corpus comprised texts from several periods, genres, and styles, including literary, scientific, and humanistic writings. For each text, the entropy per word associated to the distribution of word occurrences, h_Z, was also computed. As discussed above, the difference $D = h_Z - h$ is a measure of the increment of information when passing from a random version of the text, where each different word appears with a probability taken from Zipf's distribution, to the ordered word sequence of the actual text. In the context of information theory, D can be identified as an estimation of the Kullback–Leibler distance per word between the probability distributions of the ordered text and its random counterpart (see Sec. 1.3).

Figure 4.2 shows distributions of the values obtained for h, h_Z, and D, for the English, Finnish, and Chinese sub-corpora, comprising 5112, 101, and 392 texts, respectively. Note that the languages belong to three different families. It turns out that, for the three languages, the distributions of the entropy difference D are narrower than those of h and h_Z. Moreover, although the distributions for both h and h_Z are shifted from each other across languages, the distributions for D seem to attain their maxima around the same value.

Similar results reveal that this is also the case for the other languages considered in the study. Table 4.2 shows the average values obtained for

[2]For the study, the texts were mostly downloaded in electronic format from the public-domain depository of Project Gutenberg, www.gutenberg.org. At the time of this writing, Project Gutenberg stands as the largest free source of electronic books in various languages.

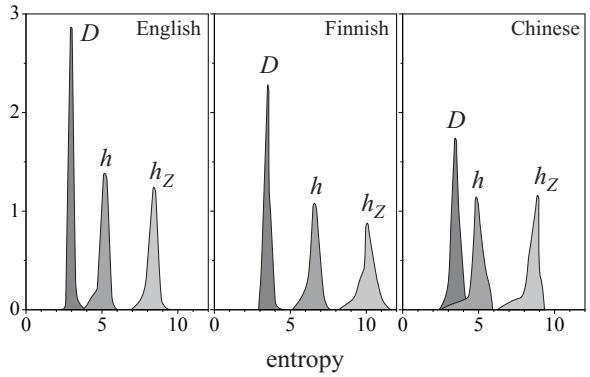

Figure 4.2: Schematic representation of the distributions of entropy per word in the actual text, h, the entropy per word in a random version of the text, h_Z, and their difference, $D = h_Z - h$, for collections of 5112 texts in English, 101 in Finnish, and 392 in Chinese. Entropies are measured in bits per word. Adapted from Montemurro and Zanette (2011).

h, h_Z, and D in each language. Over different languages, the entropies h and h_Z average, respectively, 5.3 and 8.9 bits per word. Their standard deviations, in turn, represent 23% and 14% of the respective averages. On the other hand, for the difference $D = h_Z - h$ the standard deviation is below 6% of the average of 3.6 bits per word. In other words, the value of D is much better defined across languages than the entropies h and h_Z. An equivalent conclusion stands when the Stirling approximation of Eq. (4.9) is not introduced, but the exact form of H_Z, Eq. (4.8), is used instead: for the eight languages considered in the study, the corresponding difference D is narrowly concentrated around 3.3 bits per word.

	h	h_Z	$D = h_Z - h$
English	5.7	9.2	3.5
French	5.8	9.4	3.6
German	6.2	9.8	3.5
Finnish	7.1	10.9	3.8
Tagalog	5.1	8.5	3.4
Sumerian	3.5	7.5	3.9
Old Egyptian	3.7	7.0	3.3
Chinese	5.3	9.1	3.8

Table 4.2: Entropy per word in the actual text, h, entropy per word in a random version of the text, h_Z, and their difference, D, averaged over the texts in each language. The three quantities are measured in bits per word. Adapted from Montemurro and Zanette (2011).

Due to the broad dissimilarity of grammar rules and vocabulary inflections across linguistic families, the entropies of the ordered texts and their random counterparts show substantial variations from language to language. Their difference, on the other hand, remains bounded around a well defined value. Beyond the obvious diversity between different tongues, therefore, the contribution of word ordering to the information conveyed by language emerges as a robust universal statistical feature.

It turns out that the homogeneity of the entropy related to word ordering, when compared across languages, can be understood in terms of the existence of a balance between lexical diversity and correlation lengths in word sequences (Montemurro and Zanette, 2011). Simple model languages and real human tongues show an inverse relation between word correlation lengths and the entropy per word associated with the distribution of word occurrences, h_Z. In model languages, moreover, it is possible to prove that this inverse relation is directly linked to a constant difference between h_Z and the entropy per word in the ordered text, h.

To illustrate this fact, it is useful to introduce a minimal model language, with a lexicon consisting of only two words. Word sequences are built up by means of a first-order Markov process (Gardiner, 2004) defined by the transition matrix

$$M = \begin{pmatrix} \mu_1 & 1-\mu_2 \\ 1-\mu_1 & \mu_2 \end{pmatrix}, \tag{4.12}$$

where the element M_{ij} $(i, j = 1, 2)$ is the probability that word i follows word j in the sequence. Introducing an abstract variable to identify the two words, say, $x = 0$ for the first word and $x = 1$ for the second, a correlation length λ for the word sequence can be estimated by computing the autocorrelation function (see Sec. 1.2)

$$c(\tau) = \langle x_{t+\tau} x_t \rangle - \langle x_t \rangle^2 \sim \exp\left(-\tau \ln \frac{1}{\mu_1 + \mu_2 - 1} \right), \tag{4.13}$$

namely,

$$\lambda = \left(\ln \frac{1}{\mu_1 + \mu_2 - 1} \right)^{-1}. \tag{4.14}$$

The entropy per word of the distribution of word occurrences is, in turn,

$$h_Z = -p_1 \log_2 p_1 - p_2 \log_2 p_2, \tag{4.15}$$

where $p_1 = (1-\mu_2)/(2-\mu_1-\mu_2)$ and $p_2 = (1-\mu_1)/(2-\mu_1-\mu_2)$ are the unconditional probabilities of the first and the second word, respectively. Running over different pairs of values for the parameters μ_1 and μ_2, sequences with different correlation length λ and entropy h_Z can be generated. For each ordered sequence, the entropy h, and thus the difference $D = h_Z - h$, can be estimated as well.

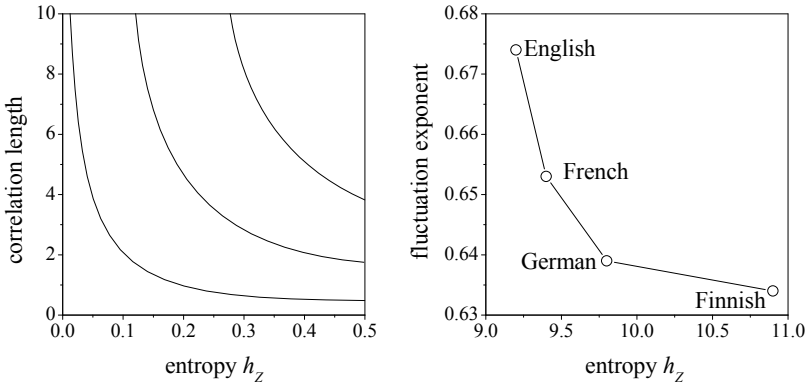

Figure 4.3: Left panel: Schematic representation of the curves of constant $D = h_Z - h$ in the plane with coordinates given by the entropy of word occurrences h_Z, measured in bits per word, and the correlation length, λ, for a Markovian two-word model language. Right panel: Fluctuation exponent γ' (which, as explained in the text, is a measure of the slowness of correlation decay) as a function of the entropy h_Z, averaged over texts in four languages. Adapted from Montemurro and Zanette (2011).

The left panel of Fig. 4.3 shows curves of constant D on the plane (h_Z, λ), for the two-word Markovian language. The inverse relation between h_Z and λ defined by these curves implies that, to maintain a fixed level for D in this class of languages, it is necessary to either increase the entropy h_Z while decreasing the correlation length λ, or *vice versa*. The same result was obtained for Markovian languages with lexicons of three and four words.

Does this relation between h_Z and the correlation length also hold for human languages? As already mentioned in Sect. 3.3, correlations in real word sequences do not decrease exponentially, as in Eq.(4.13), but rather as a power law: $c(\tau) \sim \tau^{-\gamma}$, with $\gamma > 0$ (Álvarez Lacalle *et al.*, 2006). While in this situation it is not possible to define the correlation length as a typical decay length for $c(\tau)$, the exponent γ provides, qualitatively, the same kind of measure: small and large γ correspond, respectively, to slow- and fast-decaying correlations. More directly, the fluctuation exponent $\gamma' = (2-\gamma)/2$ (Peng *et al.*, 1994; Buldyrev *et al.*, 1995) is large for long-range correlations and small for short-range correlations.

The right panel of Fig. 4.3 shows the fluctuation exponent γ' as a function of the entropy per word associated with the distribution of word occurrences, h_Z, averaged over the texts in four languages which showed statistically significant differences in γ': English, French, German, and Finnish. For each text, the exponent was calculated using the technique of detrended fluctuation analysis (Peng *et al.*, 1994). These languages, for which the difference D is approximately constant, define a curve with the same interdependency between entropy and correlation as found in the Markovian models.

As for many other quantitative parameters related to language, the diversity in the values of the entropy of word occurrences and of the word correlation length across different tongues may be ascribed to progressive divergencies along their evolution, not unlike those that separated groups of biological species from each other. The above results suggest, however, that linguistic evolutionary drift was constrained to occur keeping the difference of entropy associated with word order at a constant level. This difference seems to capture a fundamental quantitative property of language, common to a broad group of human tongues. Hence, the mechanisms underlying the way in which we assemble long words sequences to convey meaning may ultimately derive from universal cognitive constraints, inherent to the human species.

4.3 The Scales of Meaning

The entropy of word ordering, as a measure of the information needed to define a specific arrangement of words out of the mere collection of word occurrences, is a global quantitative property of the particular language stream for which it is computed. According to the discussion of Sec. 3.3, however, it is to be expected that different words contribute in different degrees to that information. Burstiness —namely, the high heterogeneity in the distribution of certain words along a text— determines a large increase in the information associated with the order of words. This contribution is granted by topical words, whose occurrences are concentrated in localized parts of the text, in contrast with functional words, whose role in language determines a more homogeneous usage.

The localization of topical words along a text is a characteristic feature of the way in which humans organize the contents of a complex message, articulating meaning as the discourse progresses. Topical words, which are borne with the semantic essence of the message, tag the parts to which they belong. From a statistical viewpoint, an interesting question is —for a long array of words conveying coherent information— to determine the optimal partition that maximizes the capability of words to tag each part. This question points to the problem of defining typical text lengths associated with the development of meaningful contents in elaborate communication (Montemurro and Zanette, 2010).

Take a text of T words, with a lexicon of V words $\{w_1, w_2, \ldots, w_V\}$, and divide it into M parts of equal size.[3] The probability that a randomly

[3]If the text length T is not an integer multiple of the number of parts M, one can truncate the text at the largest multiple of M below T. At most, the cut will be of length M for coarse partitions (small M) or T/M for fine partitions (large M). In the cases of interest in practice, which involve long texts, relevant partitions are those with $M \ll T$, and the effect of truncation is negligible.

chosen occurrence of word w_i belongs to part m is

$$p_i^m = \frac{n_i^m}{n_i}, \tag{4.16}$$

where n_i^m is the number of occurrences of w_i in that part, and n_i is its total number of occurrences all over the text. Note that, under the condition that all parts have the same size, p_i^m coincides with the homologous quantity defined in Eq. (3.16). The entropy associated with the distribution of w_i over the text partition is

$$h_i = -\sum_{m=1}^{M} p_i^m \log_2 p_i^m. \tag{4.17}$$

Being an entropy defined for each single word, the units of h_i are bits per word (see also footnote 1 in Sec. 3.3). In the present analysis, the number of parts M is a variable and, therefore, is not used as the logarithm base to normalize the entropy, in contrast with Eq. (3.17).

Much like in the study presented in the previous section, the aim here is to compare the value of h_i in an actual text with the entropy h_i^R obtained from a random shuffling of the text in question. However, there are two main differences. First, as pointed out above, h_i is now separately computed for each word w_i. Second, the degree of order of the actual text with respect to its random surrogate is measured up to the partition size, i.e. within a scale $s = T/M$. This scale characterizes the coarseness with which the distribution of words over the text is ascertained. Since n_i^m does not change by rearranging the word's occurrences within each part, h_i is not sensible to the text randomization below such scale. The average value of h_i^R over all possible word orderings can be evaluated analytically, as

$$\langle h_i^R \rangle = -M \sum_{k=1}^{K} \frac{k \begin{pmatrix} n_i \\ k \end{pmatrix} \begin{pmatrix} T - n_i \\ s - k \end{pmatrix}}{n_i \begin{pmatrix} T \\ s \end{pmatrix}} \log_2 \frac{k}{n_i}, \tag{4.18}$$

with $K = \min\{n_i, s\}$ (Montemurro and Zanette, 2010).

As an illustration of the difference between the entropy h_i and its random-text counterpart h_i^R, Fig. 4.4 shows the two quantities for the words of Charles Darwin's *On the Origin of Species*, dividing the whole text into $M = 64$ parts. To ease the comparison with each other and with $\langle h_i^R \rangle$, entropies are plotted against the number of occurrences of each word. Black and grey dots stand, respectively, for the entropies h_i and h_i^R of each individual word. The random-text entropies were calculated from a single shuffling of the original text. The curve corresponds to the average random-text entropy, given by Eq. (4.18). As expected, the random-text entropies

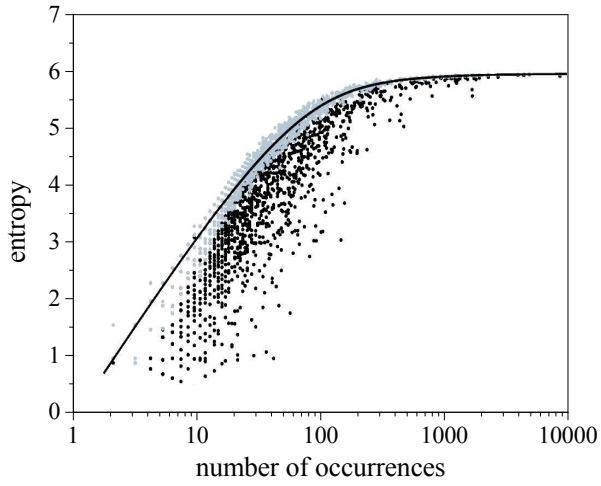

Figure 4.4: Entropies of individual words from Darwin's *On the Origin of Species*, in bits per word, plotted against the respective number of occurrences. The text was divided into 64 parts of equal length. Black and grey dots correspond, respectively, to the entropies evaluated from the ordered text, h_i, and from a random shuffling of it, h_i^R. The curve stands for the average of the random-text entropy, Eq. (4.18). Adapted from Montemurro and Zanette (2010).

are generally above those calculated from ordered version. Note that, for very large numbers of occurrences, the two entropies collapse on the same value, of $\log_2 M = 6$ bits/word. These highly frequent words —most of which are functional words such as articles, prepositions, and connectors— are distributed in a way that can hardly be discerned from a random arrangement. At the other end, for words with a few occurrences, the difference between the two entropies is also not too large, and varies around 1 bit/word. The largest discrepancies are found for words with intermediate frequencies, $10 \lesssim n_i \lesssim 200$, where the entropy difference can attain some 4 bits/word.

While, for any given word, the difference $d_i = \langle h_i^R \rangle - h_i$ is typically a positive quantity —which measures the degree of order in the word's distribution over the text partition— its specific value depends on the partition size. Indeed, consider the limiting case in which $M = 1$, so that all the word's occurrences belong to the same part, whose size equals the text length ($s = T$). In this case, any shuffling leaves the number of occurrences per part unchanged, so that $d_i = 0$. Even more, for $M = 1$ both h_i and h_i^R are identically equal to zero. Similarly, in the opposite limit where the text is divided into as many parts as words, $M = T$ ($s = 1$), each part can contain at most one occurrence. The corresponding entropies are again identical: $h_i = h_i^R = \log_2 n_i$. Thus, $d_i = 0$ also for $M = T$. On the other hand,

the entropy difference d_i should be positive for intermediate partitions. In particular, it may attain one or more maxima for $1 < M < T$.

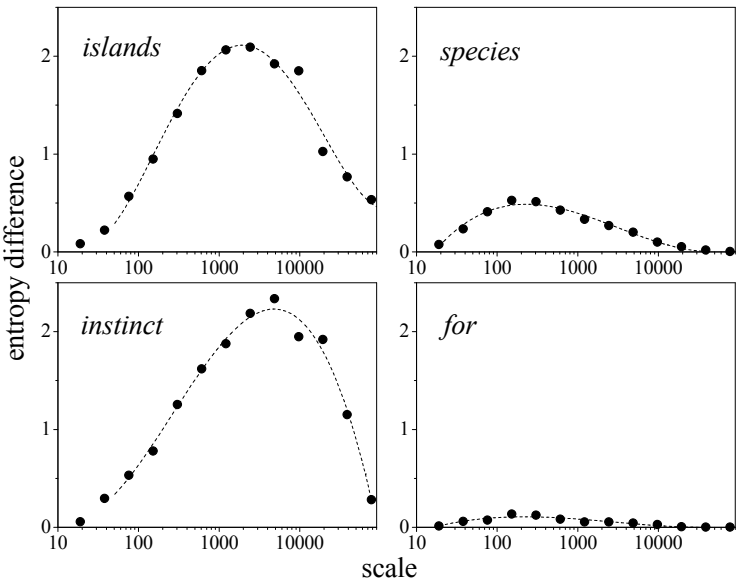

Figure 4.5: The entropy difference $d_i = \langle h_i^R \rangle - h_i$, measured in bits per word, for four words from Darwin's *On the Origin of Species*, as a function of the partition scale s. Dashed curves are polynomial fittings drawn as guides to the eye. Note that the four plots have the same scales in both axes.

Figure 4.5 shows the entropy difference d_i as a function of the scale $s = T/M$, for four words from *On the Origin of Species*: *islands*, *instinct* (see also Fig. 3.4), *species*, and *for*. In the four cases, there is a well defined intermediate maximum. However, the scale at which this maximum occurs, as well as its height, differ considerably from word to word. The two topical words *islands* and *instinct* display high maxima, both above 2 bits/word, attesting their very heterogeneous distribution along the text. The positions of the maxima —respectively, $s_{\max} \approx 2000$ and 5000— give typical scales for such heterogeneity. The word *for*, on the other hand, shows much smaller values of d_i, which discloses its very uniform use. The maximum, slightly above 0.1 bits/word, occurs for a scale $s_{\max} \approx 150$. The case of *species* is intermediate: it is a frequent word with a rather homogeneous distribution —though not as homogeneous as a typical functional word. Its maximum, at $s_{\max} \approx 200$, is above 0.5 bits/word.

The entropy differences d_i for the individual words in a text can be

integrated into a single quantity, characteristic of the whole text, given by

$$d = \sum_{i=1}^{V} f_i d_i, \tag{4.19}$$

where the sum runs over the whole lexicon, and $f_i = n_i/T$ is the total frequency of each word. The quantity d, which depends on the partition size and is also measured in bits per word, characterizes the *mutual information* between the distribution of words and the partition: the product $f_i d_i$ gauges how much information is contained in an occurrence of word w_i about its belonging to any given part of size s in the real text, with respect to a random distribution (Montemurro and Zanette, 2010). The left panel of Fig. 4.6 shows the mutual information d as a function of the scale s for three books in English: Charles Darwin's *On the Origin of Species* ($T = 155800$), Herman Meville's *Moby Dick* ($T = 218284$), and Bertrand Russell's *Analysis of the Mind* ($T = 89586$). In the three cases, the mutual information displays a well defined peak at, respectively, $s_{\max} \approx 8000$, 4000, and 3000. For this particular scale, the mutual information of the word distribution and the text partition differs maximally between the real text and a random shuffling of it. It is at this particular partition size that the distribution of words better discriminates between different sections of the text.

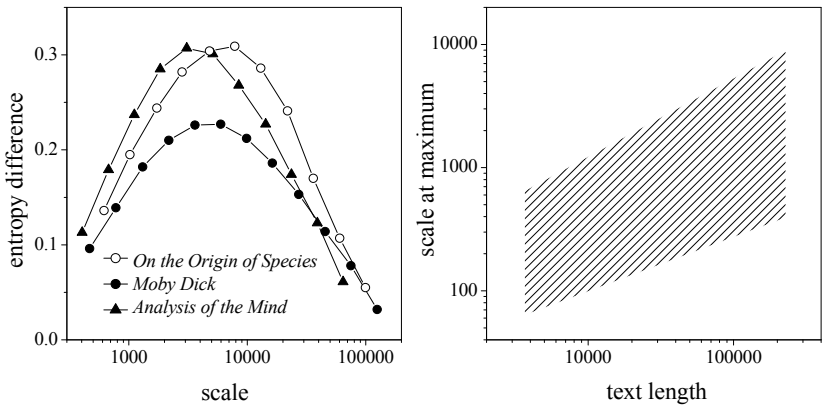

Figure 4.6: Left panel: Mutual information d, in bits per word, as a function of the partition scale s for Ch. Darwin's *On the Origin of Species*, H. Meville's *Moby Dick*, and B. Russell's *Analysis of the Mind*. Right panel: The zone in the plane of coordinates given by the text length T and the scale of maximal mutual information $s_{\max}$ occupied by a corpus of 5258 books and other writings in English. Adapted from Montemurro and Zanette (2010).

Note that, for the three books considered above, there is no direct relation between the book's length T and the scale at the maximum of the mutual information. In an analysis of a corpus containing almost 5300 texts

in English, nevertheless, it was found that there is a tendency of $s_{\max}$ to grow as T increases (Montemurro and Zanette, 2010). The corpus included English and American literature, as well as texts on science, technology and humanities. The left panel in Fig. 4.6 shows the zone in the plane $(T, s_{\max})$ where the texts of that corpus are found. It is interesting to realize that at the lower–left corner of this zone, corresponding to short texts with small $s_{\max}$, one finds collections of short quotations with no thematic unity, such as *Quotations of Lord Chesterfield* and *Quotes and Images From The Novels of Georg Ebers*. For these texts, $s_{\max}$ is below 100 words. The opposite corner, with $s_{\max} \approx 10000$, is occupied by such books as E. Gibbon's *History of The Decline and Fall of the Roman Empire* and J. Burckhardt's *Civilization of the Renaissance in Italy*, two long treatises with very consistent, uniform subjects.

Over the whole corpus, the typical scale of maximum mutual information varies around several hundred to a few thousand words. Clearly, such lengths are far beyond the scope of the rules of syntax. Syntactic constraints, in fact, apply at the level of phrases and sentences, spanning several words at most. The values of $s_{\max}$, on the other hand, should be related to the characteristic scales associated with the development of the successive thematic subjects covered by a book. If the text is divided into sections of size $s_{\max}$, the distribution of words in those parts bears maximum information with respect to a random distribution. For single texts with overall semantic unity, $s_{\max}$ should represent the typical length of the spans over which semantic structures develop.

On the Origin of Species	*Moby Dick*	*Analysis of the Mind*
species	*I*	*image*
varieties	*whale*	*memory*
hybrids	*you*	*word*
forms	*Ahab*	*belief*
islands	*is*	*desire*
selection	*Queequeg*	*sensations*
genera	*thou*	*object*
plants	*he*	*past*
seeds	*captain*	*knowledge*
sterility	*boat*	*contents*

Table 4.3: Some of the words with largest individual contribution to the total mutual information d, in three books in English. Adapted from Montemurro and Zanette (2010).

This conjecture is supported by the identification, in each text, of the words whose contribution to the overall mutual information d —given, according to Eq. (4.19), by the product $f_i d_i$— is large. Table 4.3 extracts

some of the words with the largest contributions to d in *On the Origin of Species*, *Moby Dick*, and *Analysis of the Mind*. Their relation to the semantic contents of each book is apparent in Darwin's and Russell's treatises. In Meville's novel, such words as the pronouns *I*, *you*, and *he* —which in other texts would play an essentially functional role— attain high relevance due to the narrative style of the work, whose structure is built up around its characters.

The typical scales $s_{\max}$ which emerge from this quantitative analysis coincide with the text lengths that, in English, are widely recognized as the span which allows for the development of a coherent line of argument. As an example, at the time of the present writing (August 2011), the influential scientific journal *Nature* was accepting manuscripts of two kinds, Letters and Articles, whose recommended lengths were 1500 and 3000 words, respectively. Editorial comments in the same journal, in turn, contained some 600 words. The magazine *Science* accepted longer articles, but all shorter contributions —Brevia, Technical Comments, Perspectives, among others— were limited to around 1000 words in length.

The analysis of the above mentioned corpus included also the determination of the maximal mutual information per word, $d_{\max}$, in each book. It turns out that the distribution of $d_{\max}$ over the whole corpus is sharply peaked around 0.2 bits/word. Therefore, whereas the partition size which maximizes the mutual information, $s_{\max}$, varies substantially between different texts in English, the accuracy with which words tag each part is much more uniform.

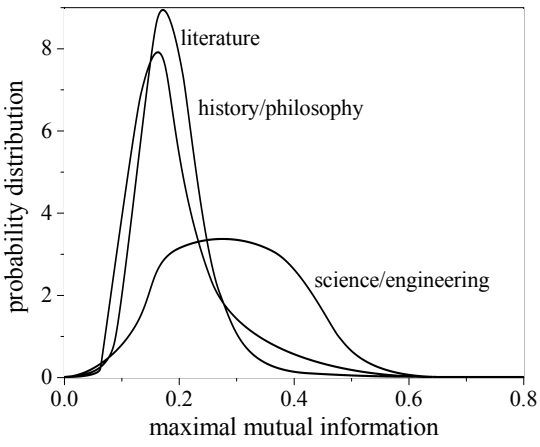

Figure 4.7: Distribution of the maximal mutual information per word, $d_{\max}$, measured in bits per word, for three sub-corpora of English books: 3329 literary words, 1374 books on history and philosophy, and 555 books on science and engineering. Adapted from Montemurro and Zanette (2010).

This picture changes, however, when books are differentiated from each other by their main theme or genre. The corpus can be divided into three groups: 3329 literary texts, 1374 treatises on history and philosophy, and 555 books on science and engineering. Figure 4.7 shows the distribution of the mutual information at the maximum, $d_{\max}$, over the three groups. It turns out that literary and humanistic texts tend to have lower values of the information per word than the books on science and engineering. The latter have a broader distribution, extending over higher values of the maximal mutual information. This indicates that the usage of language in scientific and engineering books is such that the distribution of words tags more efficiently the different parts of the text.

In summary, the study of the mutual information of the distribution of individual words, with respect to a given partition of the text, confirms that, as seen in Sec. 3.3, heterogeneity and topicality are closely related features. Additionally, we see now that the information-theoretical approach to word distributions makes it possible to identify, for each text, a typical scale given by the partition size at which the mutual information attains a maximum. For such partition, word occurrences tag most efficiently the different parts of the text. The analysis of a large corpus of English books showed that those typical scales vary between several hundred to a few thousand words, and are therefore much longer than the range at which grammatical rules apply. They are rather related to the spans needed to expound and unfold the contents that an elaborate and coherent piece of human language conveys.

4.4 Word Patterns Across Texts

The same kind of information-theoretical techniques discussed in Secs. 3.3 and 4.3 for analyzing the structures of the distribution of words among different parts of a given text, can be used to disclose patterns of word usage across different texts. In a corpus formed by the concatenation of several texts —each of them being a self-consistent piece of language, with its own style and topic— heterogeneity in word usage may emerge from a variety of factors. Very topical words, related to the specific subjects of just one or a few texts in the corpus, will be highly concentrated in those parts. The entropy associated with their distribution will correspondingly be relatively small. On the other hand, functional words with no special relation to particular topics should display a more homogeneous distribution. However, their marginal heterogeneities may be the signature of more subtle differences. Genre and style, indeed, are expected to have an effect on the usage of such words, influencing both their frequency and their arrangement along each text. The quantitative analysis of differences in the usage of the most frequent words —including articles, prepositions, pronouns, and connectors, which have a mainly functional role in language— is in fact a

basic tool in authorship attribution (Burrows, 1987, 1992).

Heterogeneity in the distribution of word occurrences has been studied in connection with the grammatical role of each word, in a corpus formed by the concatenation of 36 plays by William Shakespeare (Montemurro and Zanette, 2002a), including tragedies and comedies of historical (Roman and English) and fictional inspiration. The average length of the plays was close to 24600 words, and the corpus had 23150 different words. For each word w_i, the entropy h_i given by Eq. (3.17) was calculated, with each part of the corpus corresponding to an individual play ($M = 36$). By construction, $h_i = 0$ for words that occur in only one play, and $h_i \approx 1$ for highly frequent, evenly distributed words. If all plays had equal lengths, an evenly distributed word with a total of n_i occurrences —for which the number of occurrences in each play were approximately proportional to the play's length— would have $h_i \approx 1 - (M-1)/2n_i \ln M$ for large n_i.

The upper panel of Fig. 4.8 shows the quantity $(1-h_i)n_i$ for all the words in the Shakespeare corpus plotted against their number of occurrences n_i. Words over the oblique dashed line are those with $h_i = 0$. The horizontal dashed line, whose vertical coordinate equals $(M-1)/2 \ln M$, would in turn coincide with the position of very frequent words with a homogeneous distribution over a partition of the corpus into equal parts. It turns out that the words in this corpus are widely spread between the two boundaries. The identification of different word types in different parts of this plot confirms some expected features, but also discloses less obvious patterns.

Predictably, close to the line $h_i = 0$ one finds proper nouns corresponding to character names —many of which are unique to a single play— but also to the name of countries, towns, and other places. Each of these words is strongly related to, at most, a few parts of the corpus. Slightly less specific to individual plays are nouns who refer to the status of human beings, in particular, to nobility titles —which are highly frequent, and topically very important, in Shakespeare's plays. The position of some of these words has been identified in the upper panel of Fig. 4.8.

The analysis of the Shakespeare corpus comprised the classification of the 2000 most frequent words into five classes, according to their grammatical function (Montemurro and Zanette, 2002a). This sub-lexicon coincides, approximately, with all the words that occur more than 40 times in the corpus. Two of the classes were, as mentioned above, proper nouns, and nouns referring to nobility status. The other three classes were pronouns, verbs and adverbs, and common nouns and adjectives. The lower panel of Fig. 4.8 shows the regions of the same plane of the upper panel which are occupied by each class.

As advanced, proper and nobility nouns (respectively, zones A and B) lie along the line $h_i = 0$, the former closer than the latter. The case of pronouns (zone C) is more striking. While, not unexpectedly, they are located in the region corresponding to very frequent words, with large n_i, their relatively

high position in the plot reveals a surprising degree of heterogeneity across the corpus. Since all Shakespeare's plays are supposed to have been written by the same author, within consistent genre and style, an uneven use of pronouns —which have a mainly functional linguistic role— seems anomalous. An account of this feature in terms of the plays' themes, which are their more conspicuous difference with each other, seems unlikely. Until now, the anomaly has not been satisfactorily explained or explored further.

A comparison of verbs and adverbs (zone D) with common nouns and adjectives (zone E) is also interesting. While the two zones intersect in a wide region, the average position of common nouns and adjectives is clearly above that of verbs and adverbs, revealing that the latter are more homogeneously distributed over the corpus than the former. This difference, again, has not been explained. Is it that, because of the very nature of the theatrical genre, objects are more specific to a given subject than the actions in which they are involved? Or, perhaps, is this the consequence of the essentially different semantic nature of nouns and verbs —which, however, appear so closely linked in the communication of any meaningful message? These questions are still open. Answering them requires, along with sound linguistic argumentation, extensive statistical analysis of large corpora in different tongues, genres, and styles which, in spite of some consistent efforts in the last few decades, remains largely incipient.

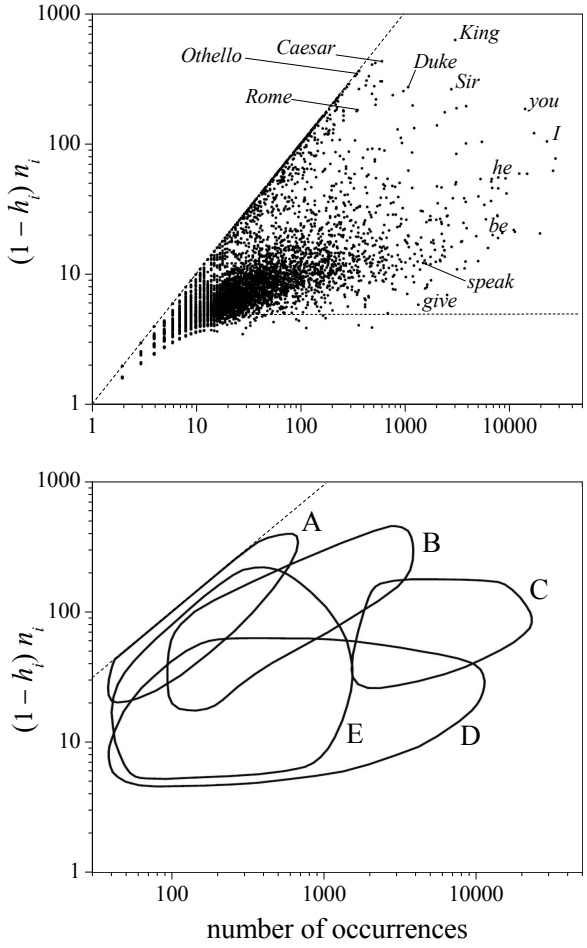

Figure 4.8: Upper panel: The quantity $(1 - h_i)n_i$ as a function of the number of occurrences n_i for the 23150 different words of 36 plays by William Shakespeare. The entropy h_i is measured in bits per word. The oblique dashed line corresponds to $h_i = 0$. The horizontal dashed line is the value expected for frequent, evenly distributed words, in the case that all the plays has the same length. The position of several words is identified by labels. Lower panel: Schematic representation of the zones occupied by five grammatical classes over the same plane as in the upper panel: (A) proper nouns, (B) nouns referring to nobility status, (C) pronouns, (D) verbs and adverbs, and (E) common nouns and adjectives. Note that the horizontal scales of the two panels are not the same. Adapted from Montemurro and Zanette (2002a).

Bibliography

Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Dover, New York).

Álvarez Lacalle, E., Dorow, B., Eckmann, J. P. and Moses, E. (2006). Hierarchical structures induce long-range dynamical correlations in written texts, *Proc. Natl. Acad. Sci. USA* **103**, 21, pp. 7956–7961.

Applebaum, D. (2005). *Lévy Processes and Stochastic Calculus* (Cambridge University, Cambridge).

Auerbach, F. (1913). Das Gesetz der Bevölkerungskonzentration, *Petermans Mitteilungen* **59**, 1, pp. 74–76.

Bernstein, L. (1973). *The Unanswered Question* (Harvard University, Cambridge).

Boroda, M. G. and Polikarpov, A. A. (1988). The Zipf–Mandelbrot law and units of different text levels, *Musikometrika* **1**, 1, pp. 127–158.

Borovik, A. S., Grosberg, A. Y. and Frank Kamenetskii, M. D. (1994). Fractality of DNA texts, *J. Biomolec. Struct. Dyn.* **12**, 3, pp. 655–669.

Brown, C. M. and Hagoort, P. (2000). *The Neurocognition of Language* (Oxford University, Oxford).

Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Matsa, M. E., Peng, C.-K., Simons, M. and Stanley, H. E. (1995). Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis, *Phys. Rev. E* **51**, 5, pp. 5084–5091.

Burlando, B. (1990). The fractal dimension of taxonomic systems, *J. Theor. Biol.* **146**, 1, pp. 99–114.

Burlando, B. (1993). The fractal geometry of evolution, *J. Theor. Biol.* **163**, 2, pp. 161–172.

Burrows, J. F. (1987). Word patterns and story shapes: The statistical analysis of narrative style, *Liter. Ling. Comput.* **2**, 1, pp. 61–70.

Burrows, J. F. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information, *Liter. Ling. Comput.* **7**, 2, pp. 91–109.

Church, K. W. and Gale, W. A. (1995). Poisson mixtures, *Nat. Lang. Eng.* **1**, 2, pp. 163–190.

Consul, P. C. (1991). Evolution of surnames, *Int. Stat. Rev.* **59**, 3, pp. 271–278.

Cover, T. M. and King, R. C. (1978). Convergent gambling estimate of entropy of English, *IEEE Trans. Inf. Theor.* **24**, 4, pp. 413–421.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory* (Wiley-Interscience, Hoboken).

Ebeling, W. and Neiman, A. (1995). Long-range correlations between letters and sentences in texts, *Physica A* **215**, 3, pp. 233–241.

Ebeling, W. and Poschel, T. (1994). Entropy and long-range correlations in literary English, *Europhys. Lett.* **26**, 4, pp. 241–246.

Ehrenfest, P. and Ehrenfest, T. (1911). *Begriffliche Grundlagen der statistischen Auffassung in der Mechanik*, *Enzyklopädie der mathematischen Wissenschaften mit Einschluss ihrer Anwendungen*, Vol. IV.2 (Teubner, Leipzig), pp. 3–90.

Eldridge, R. C. (1911). *Six Thousand Common English Words* (Clement, Buffalo).

Feder, J. (1988). *Fractals* (Plenum, New York).

Ferrer i Cancho, R. and Elvevåg, R. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution, *PLoS ONE* **5**, 3, p. e9411.

Ferrer i Cancho, R. and Solé, R. V. (2001a). Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited, *J. Quant. Linguistics* **8**, 3, pp. 165–173.

Ferrer i Cancho, R. and Solé, R. V. (2001b). The small-world of human language, *Proc. Roy. Soc. London B* **268**, 1482, pp. 2261–2265.

Ferrer i Cancho, R. and Solé, R. V. (2003). Least effort and the origins of scaling in human language, *Proc. Natl. Acad. Sci. USA* **100**, 3, pp. 788–791.

Gao, Y., Kontoyiannis, I. and Bienenstock, E. (2008). Estimating the entropy of binary time series: Methodology, some theory and a simulation study, *Entropy* **10**, 1, pp. 71–99.

Gardiner, C. (2004). *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences* (Springer, Berlin).

Gelbukh, A. and Sidorov, G. (2001). *Zipf and Heaps laws' coefficients depend on language*, *Lecture Notes in Computer Science*, Vol. 2004 (Springer, Berlin), pp. 332–335.

Gibrat, R. (1932). *Les inégalités économiques* (Sirey, Paris).

Greenberg, J. H. (1963). *Universals of Languages* (MIT Press, Cambridge).

Hanley, M. L. (1937). *Word Index to James Joyce's Ulysses* (University of Wiconsin, Madison).

Harris, T. E. (1963). *The Theory of Branching Processes* (Springer, Berlin).

Heaps, H. S. (1978). *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, Orlando).

Herrera, J. P. and Pury, P. A. (2008). Statistical keyword detection in literary corpora, *Eur. Phys. J. B* **63**, 1, pp. 135–146.

Huang, K. (1987). *Statistical Mechanics* (Wiley, New York).

Indurkhya, N. and Damerau, F. J. (2010). *Handbook of Natural Language Processing* (Chapman and Hall–CRC, Boca Raton).

Islam, M. N. (1995). A stochastic model for surname evolution, *Biom. J.* **37**, 1, pp. 119–126.

Kanter, I. and Kessler, D. A. (1995). Markov processes: Linguistics and Zipf's law, *Phys. Rev. Lett.* **74**, 22, pp. 4559–4562.

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling, *Nat. Lang. Eng.* **2**, 1, pp. 15–59.

Klug, W. S. and Cummings, M. R. (1997). *Concepts of Genetics* (Prentice–Hall, Upper Saddle River).

Kontoyiannis, I., Algoet, P. H., Suhov, Y. M. and Wyner, A. J. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to English text, *IEEE Trans. Inf. Theor.* **44**, 12, pp. 1319–1327.

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch* (Oxford University, Oxford).

Landini, G. (2001). Evidence of linguistic structure in the Voynich Manuscript using spectral analysis, *Cryptologia* **25**, 4, pp. 275–295.

Lerdahl, F. and Jackendoff, R. (1983). *A Generative Theory of Tonal Music* (MIT Press, Cambridge).

Lewis, M. P. (2009). *Ethnologue: Languages of the World* (SIL International, Dallas).

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution, *IEEE Trans. Inform. Theor.* **38**, 6, pp. 1842–1845.

Lü, L., K.Zhang, Z. and Zhou, T. (2010). Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems, *PLoS ONE* **5**, 12, p. e14139.

Maess, B., Koelsch, S., Gunter, T. and Friederici, A. D. (2001). Musical syntax is processed in Brocas area: an MEG study, *Nature Neurosci.* **4**, pp. 540–545.

Manaris, B., Vaughan, D., Wagner, C., Romero, J. and Davis, R. B. (2003). *Evolutionary music and the Zipf–Mandelbrot law: Progress towards developing fitness functions for pleasant music*, *Lecture Notes in Computer Science*, Vol. 2611 (Springer, Berlin), pp. 522–534.

Mandelbrot, B. B. (1951). Adaptation d'un message à la ligne de transmission. I & II, *Comptes Rendues (Paris)* **232**, pp. 1638–1740 & 2003–2005.

Mandelbrot, B. B. (1955). *On recurrent noise limiting coding* (Interscience, New York), pp. 205–221.

Mandelbrot, B. B. (1959). A note on a class of skew distribution function. Analysis and critique of a paper by H. A. Simon, *Information and Control* **2**, 1, pp. 90–99.

Mandelbrot, B. B. (1997). *Fractals and Scaling in Finance. Discontinuity, Concentration, Risk* (Springer, New York).

Manrubia, S. C., Derrida, B. and Zanette, D. H. (2003). Genealogy in the era of genomics, *American Scientist* **41**, 2, pp. 158–165.

Manrubia, S. C. and Zanette, D. H. (2001). Vertical transmission of culture and the distribution of family names, *Physica A* **295**, 1, pp. 1–8.

Manrubia, S. C. and Zanette, D. H. (2002). At the boundary between biological and cultural evolution: The origin of surname distributions, *J. Theor. Biol.* **216**, 4, pp. 461–477.

Miller, G. A. (1965). *Introduction*, to G. K. Zipf, *The Psycho-Biology of Language. An Introduction to Dynamic Philology* (MIT Press, Cambridge, 1965), pp. v–x.

Miller, G. A. (1981). *Language and Speech* (Freeman, San Francisco).

Miyazima, S., Lee, Y., Nagamine, T. and Miyajima, H. (2000). Power-law distribution of family names in Japanese societies, *Physica A* **278**, 1–2, pp. 282–288.

Montemurro, M. A. (2001). Beyond the Zipf–Mandelbrot law in quantitative linguistics, *Physica A* **300**, 3–4, pp. 567–578.

Montemurro, M. A. and Pury, P. A. (2002). Long-range fractal correlations in literary corpora, *Fractals* **10**, 4, pp. 451–461.

Montemurro, M. A. and Zanette, D. H. (2002a). Entropic analysis of the role of words in literary texts, *Adv. Compl. Sys.* **5**, 1, pp. 7–17.

Montemurro, M. A. and Zanette, D. H. (2002b). New perspectives on Zipf's law: from single texts to large corpora, *Glottometrics* **4**, 1, pp. 86–98.

Montemurro, M. A. and Zanette, D. H. (2009). The statistics of meaning: Darwin, Gibbon and Moby Dick, *Significance* **6**, 4, pp. 165–169.

Montemurro, M. A. and Zanette, D. H. (2010). Towards the quantificantion of the semantic information encoded in written language, *Adv. Compl. Sys.* **13**, 2, pp. 135–153.

Montemurro, M. A. and Zanette, D. H. (2011). Universal entropy of word ordering accross linguistic families, *PLoS ONE* **6**, 5, p. e19875.

Murray, J. D. (2002). *Mathematical Biology. An Introduction* (Springer, New York).

Ortuño, M., Carpena, P., Bernaola Galván, P., Muñoz, E. and Somoza, A. M. (2002). Keyword detection in natural languages and DNA, *Europhys. Lett.* **57**, 5, pp. 759–764.

Panaretos, J. (1989). On the evolution of surnames, *Int. Stat. Rev.* **57**, 2, pp. 161–167.

Pareto, V. (1896). *Cours d'économie politique* (F. Rouge, Lausanne).

Patel, A. D. (2008). *Music, Language, and the Brain* (Oxford University, New York).

Peng, C.-K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. and Stanley, H. E. (1992). Long-range correlations in nucleotide sequences, *Nature* **356**, 6365, pp. 168–170.

Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E. and Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides, *Phys. Rev. E* **49**, 2, pp. 1685–1689.

Schenkel, A., Zhang, J. and Zhang, Y. (1993). Long-range correlations in human writings, *Fractals* **1**, 1, pp. 47–57.

Schwartz, B. and Reisberg, D. (1991). *Learning and Memory* (W. W. Norton and Co., New York).

Shannon, C. E. (1948a). A mathematical theory of communication, *Bell Syst. Tech. J.* **27**, 3, pp. 379–423.

Shannon, C. E. (1948b). A mathematical theory of communication — Part III, *Bell Syst. Tech. J.* **27**, 4, pp. 623–656.

Shannon, C. E. (1951). Prediction and entropy of printed English, *Bell Syst. Tech. J.* **30**, 1, pp. 50–64.

Simon, H. A. (1955). On a class of skew distribution functions, *Biometrika* **42**, 3–4, pp. 425–440.

Simon, H. A. (1957). *Models of Man. Social and Rational* (Wiley, New York).

Simon, H. A. (1961). Reply to Dr. Mandelbrot's post scriptum, *Information and Control* **4**, 3, pp. 305–308.

Sornette, D. (1998). Multiplicative processes and power laws, *Phys. Rev. E* **57**, 4, pp. 4811–4813.

Sornette, D. (2000). *Critical Phenomena in Natural Sciences. Chaos, Fractals, Seforganization, and Disorder: Concepts and Tools* (Springer, Berlin).

Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Goldberger, Z. D., Havlin, S., Mantegna, R. N., Ossadnik, S. M., Peng, C. K. and Simons, M. (1994). Statistical mechanics in biology: how ubiquitous are long-range correlations? *Physica A* **205**, 1–3, pp. 214–253.

Teahan, W. J. and Cleary, J. G. (1996). The entropy of English using PPM-based models, in *Proc. Data Compression Conference (DCC'96)* (Snowbird, USA), pp. 53–62.

van Eemeren, F. H. (2001). *Crucial Concepts in Argumentation Theory* (Univerity of Chicago, Chicago).

Voss, R. F. (1985). *Random fractal forgeries* (Plenum, New York), pp. 1–11.

Willis, J. C. and Yule, G. U. (1922). Some statistics of evolution and geographical distribution in plants and animals, and their significance, *Nature* **109**, pp. 177–179.

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, *Phyl. Trans. R. Soc. London Ser. B* **213**, pp. 21–87.

Zanette, D. H. (2001). Self-similarity in the taxonomic classification of human languages, *Adv. Compl. Sys.* **4**, 2–3, pp. 281–286.

Zanette, D. H. (2006). Zipf's law and the creation of musical context, *Musicae Scientiae* **10**, pp. 3–18.

Zanette, D. H. (2008). Playing by numbers, *Nature* **453**, pp. 988–989.

Zanette, D. H. and Manrubia, S. C. (2007). *Multiplicative processes in social systems*, *World Scientific Lecture Notes in Complex Systems*, Vol. 7, chap. 6 (World Scientific, Singapore), pp. 129–158.

Zanette, D. H. and Montemurro, M. A. (2005). Dynamics of text generation with realistic Zipf's distribution, *J. Quant. Linguistics* **12**, 1, pp. 29–40.

Zhou, H. and Slater, G. W. (2003). A metric to search for relevant words, *Physica A* **329**, 1–2, pp. 309–327.

Zipf, G. K. (1936). *The Psycho-Biology of Language. An Introduction to Dynamic Philology* (Routledge, London).

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort. An Introduction to Human Ecology* (Addison–Wesley, Cambridge).

Ziv, J. and Lempel, A. (1977). Universal algorithm for sequential data compression, *IEEE Trans. Inf. Theor.* **23**, 4, pp. 337–343.