

STATISTICAL KEYWORD DETECTION IN LITERARY CORPORA

JUAN P. HERRERA* and PEDRO A. PURY†

*Facultad de Matemática,
Astronomía y Física,
Universidad Nacional de Córdoba,
Ciudad Universitaria,
X5000HUA Córdoba, Argentina*

Abstract

Understanding the complexity of human language requires an appropriate analysis of the statistical distribution of words in texts. We consider the information retrieval problem of detecting and ranking the relevant words of a text by means of statistical information referring to the *spatial* use of the words. The Shannon's entropy of information is used as tool for automatic keyword extraction. Using *The Origin of Species* by Charles Darwin as prototypical example, we show the performance of our detector and compare it with another proposals in the literature. Particular emphasis is addressed to the random texts for calibrating the ranking indices.

PACS numbers: 89.70.+c, 05.45.Tp, 89.75.-k

Keywords: Text mining; word clustering; random texts; Shannon entropy.

* Present address: Intel Software Development Center (e-mail: jherrera14@gmail.com).

† Corresponding author (e-mail: pury@famaf.unc.edu.ar).

I. INTRODUCTION

Data mining for texts is a well-established area of natural language processing [1]. Text mining is the computerised extraction of useful answers from a mass of textual information by machine methods, computer-assisted human ones, or a combination of the two. A key problem in text mining is the extraction of keywords from texts for which no *a priori* information is available. The problem of unsupervised extraction of relevant words from their statistical properties was first addressed by Luhn [2]. He based his method in the analysis of frequencies of Zipf [3]. The Zipf's analysis consists of counting the number of occurrences of each distinct word in a given text, and then generating a list of all these words ordered by decreasing frequency. In this list, each word is identified by its position or *Zipf's rank* in the list. The empirical observation of Zipf was that the frequency of occurrence of r -th rank in the list is proportional to r^{-1} (*Zipf's law*). Luhn proposed the crude approach of exclude the words in both ends of the Zipf's list and consider as keywords the remaining cases. The limitations of the Luhn's approach are known in the literature [4].

The main goal of this work is to research in the unsupervised statistical methods for detecting keywords in literacy texts beyond the simple counting of word occurrences. We are searching for relevance according to the text's context, but we will only use statistical information about the *spatial* use of the words in a text. Particularly, the measure of content of information of each word can be made by the Shannon's entropy. In the physics literature we can found several applications of the entropy concept to linguistics and natural language like DNA sequences analysis [5, 6, 7], measure of long-range correlations [8, 9], language acquisition [10], authorship disputes [11, 12], communication model [13], and linguistic parsing [14]. For our task, in Sec. II we first review the algorithms proposed in the literature based on the analysis of statistical distribution of words in a text. Then, in Sec. III we discuss the behaviour of the indices in random texts. Using the Shannon's entropy, in Sec. IV we develop another index based on the information content of the sequence of occurrences of each word in the text.

We will use as prototypical real text “*On the Origin of Species by Means of Natural Selection, or The Preservation of Favoured Races in the Struggle for Life*” [15] (usually abbreviated to *The Origin of Species*) by Charles Darwin (1859). The book was written with the vocabulary of a nineteenth-century naturalist but with the fluid prose of a popular

author, that combines narration of journeys with academic descriptions. Thus, technical terms (like INSTINCT) are scattered in Darwin’s observations from his voyage on the Beagle. For the preparation of our working corpus we first withdrew any punctuation symbol from the text, mapped all words to uppercase and then used the simple tokenization method based on whitespaces. We draw a distinction between a word token versus a word type. For our convenience, we define a word type as any different string of letters between two spaces. Thus, for our elementary analysis, words like INSTINCT and INSTINCTS correspond to different word types in our corpus. On the other hand, a word token is each individual occurrence of a given word type. The relevant words have not been explicitly defined in the Darwin’s book, with exception of a glossary appended at the end of the work. Therefore, the table of contents in the beginning, the glossary and the analytical index, also inserted at the end, were removed from our corpus. In this manner, we avoid introducing obvious bias for the words used in these parts. Thus, the prepared corpus has got 94% of material from the original Darwin’s book and contains 192665 word tokens and 8294 word types. The corpus has got 842 paragraphs distributed in 16 chapters. From the deleted sections of the Darwin’s book we selected by hand a set of keywords as glossary. In sec. V we use it for measuring the performace of each index as keyword’s detector. Finally, in Sec. VI we present a summary of the work. The mathematical notes are relegated to the appendices. In A we review the geometrical distribution, useful to random texts, and in B we calculate the entropy of a random text.

II. CLUSTERING AS CRITERION FOR RELEVANCE OF WORDS

The attraction between words is a phenomenon that play an important role in both language processing and acquisition, and it was modeled for information retrieval and speech recognition purposes [16, 17]. Empirical data reveals that the attraction between words decays exponentially, while stylistic and syntactic constraints create a repulsion between words that discourages close co-occurrence. The attraction-repulsion phenomenon is more accentuated for relevant words than for common words, which have got less syntactic penalties for close co-occurrence. Therefore, the spatial distributions of relevant words in the text are inhomogeneous and these words gather together in some portions of the text forming clusters. The clustering phenomenon can be visualised in Fig. 1 where we have plotted the

absolute positions of four different word types from the Darwin’s corpus in a “bar code” arrangement. The clustering becomes manifest in the patterns of NATURAL, LIFE, and INSTINCT in spite of their different numbers of occurrences; whereas is completely absent from the “bar code” of THE, which is the more frequent word in the English language.

Recently, the assumption that highly relevant words should be concentrate in some portions of the text was used to searching relevant words in a given text. In the following two subsection, we briefly review the indices of relevance of words proposed by Ortuño et al. [18] and Zhou and Slater [19], which are based on the spatial distribution of words in the text.

A. σ -index

To study the spatial distribution of a given word type in a text, we can map the occurrences of the corresponding word tokens into a time series. For this task, we denote by t_i the absolute position in the corpus of the i -th occurrence of the token. Thus, we obtain the sequence $\{t_0, t_1, \dots, t_n, t_{n+1}\}$, where we are assuming that there are n word tokens. We have additionally included the boundaries of the corpus defining $t_0 = 0$ and $t_{n+1} = N + 1$, where N is the total number of tokens in the corpus, in order of take into account the space before the first occurrence and the space after the last occurrence [19].

Given the sequence of token’s occurrences $\{t_0, t_1, \dots, t_n, t_{n+1}\}$, the average distance between two successive tokens is given by

$$\mu = \frac{1}{n+1} \sum_{i=0}^n (t_{i+1} - t_i) = \frac{N+1}{n+1}, \quad (1)$$

and the sample standard deviation of the set of spacings between nearest-neighbour word tokens $(t_{i+1} - t_i)$ is by definition

$$S = \sqrt{\frac{1}{n} \sum_{i=0}^n ((t_{i+1} - t_i) - \mu)^2}. \quad (2)$$

To eliminate the dependence on the frequency of occurrence (n/N) for different word types, in Ref. [18] the authors suggest to normalise the token spacings, *i.e.*, to measure them in units of their corresponding mean value. Thus, we define

$$\sigma = \frac{S}{\mu}. \quad (3)$$

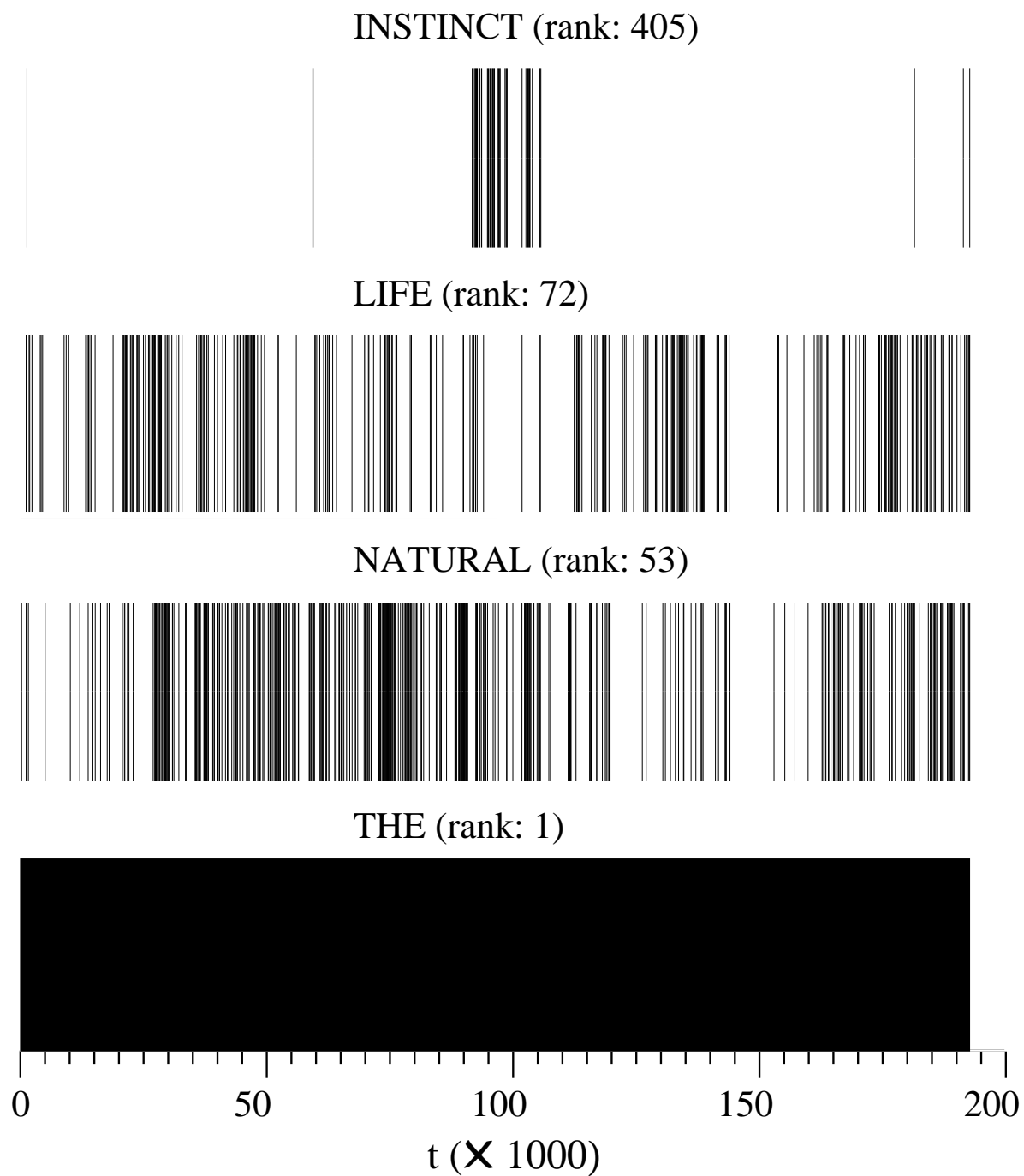


FIG. 1: Absolute positions in the corpus based on *The Origin of Species* by Charles Darwin, of the word types: THE (13414 occurrences), NATURAL (475 occurrences), LIFE (326 occurrences), and INSTINCT (64 occurrences).

Given that the standard deviation grows rapidly when the inhomogeneity of the distribution of spacing $t_{i+1} - t_i$, increases, Ortuño et al. [18] proposed σ as an indicator of the relevance of the words in the analysed text. In many cases, empirical evidence vindicates that large σ values generally correspond to terms relevant to the considered text, and that common words have associated low values of σ . However, Zhou and Slater [19] pointed out that σ -index has some weaknesses. First, several obviously common (relevant) words have relative high (low) σ values in several texts. Second, the index is not stable in the sense that can be strongly affected by the change of a single occurrence position. Third, high values of σ do not always imply a cluster concentration. A big word cluster can be break into small clusters without substantial change in the σ value.

B. Γ -index

σ -index only compute the spacings between nearest-neighbour word tokens. To improve the performance in the searching for relevance, Zhou and Slater [19] introduced a new index that use more information from the sequence of occurrences $\{t_0, t_1, \dots, t_n, t_{n+1}\}$. For this task, these authors consider the spacings $w_i = t_i - t_{i-1}$, with $i = 1, \dots, n+1$, and define the *average separation* around the occurrence at t_i as

$$d(t_i) = \frac{w_{i+1} + w_i}{2} = \frac{t_{i+1} - t_{i-1}}{2}, \quad i = 1, \dots, n. \quad (4)$$

The position t_i is said to be a cluster point if $d(t_i) < \mu$. The new suggestion is that the relevance of a word in a given text is related to the number of cluster points found in it. Thus, in order to measure the degree of clusterization, the local *cluster index* at position t_i is defined by

$$\gamma(t_i) = \begin{cases} \frac{\mu - d(t_i)}{\mu} & \text{if } t_i \text{ is a cluster point} \\ 0 & \text{otherwise} \end{cases}. \quad (5)$$

Finally, a new index for measure relevance is obtained from the average of all *cluster indices* corresponding to a given word type

$$\Gamma = \frac{1}{n} \sum_{i=1}^n \gamma(t_i). \quad (6)$$

Γ -index is more stable than σ , but is still based on local information of occurrences and computationally more complex to evaluate.

III. RANDOM TEXT AND SHUFFLED WORDS

In a text completely at random, a word type w is only characterised by their frequency of occurrence (p_w). Thus, a random text can be generated by picking successively tokens by chance in such a way that at each position the probability of find a token, corresponding to the word type w , is p_w . Obviously, $\sum_w p_w = 1$. For the word type w , we have in this manner defined a binomial experiment where the probability of success (occurrence) at each site in the text is p_w , and the probability of failure (non-occurrence) is $(1 - p_w)$. Therefore, the distribution of distances between nearest neighbour tokens corresponding to the same word type is geometrical. In A we have compiled some results of the geometrical distribution that are useful for our next analyses.

Besides its worth as comparative standard, the theoretical random text has the virtue of being analytically tractable. As well, from an empirical point of view, there is a workable fashion for building a random version of a corpora. In an actual corpora the probabilities of occurrence p are estimated from the frequencies n/N , where n is the number of tokens corresponding to a given word type and N is the total number of tokens in the corpus. A random version of the text can be obtained shuffling or permuting all the tokens. The random shuffling of all the words has the effect of rescasting the corpus into a nonsensical realization, keeping the same original tokens without discernible order at any level. However, both the Zipf's list of ranks and the frequency of occurrence of each word type are kept intact.

The important point that we want to stress is that the indices of relevance defined in the previous section are functions of the frequencies of occurrence of each word type. Thus, in a random text the values of these indices change with p , which has nonsense. In a truly random text, there are not relevant words. Therefore, to eliminate completely the dependence on frequency we need to renormalise the indices with their values in the random version of the corpus.

A. Renormalised σ -index

For a given probability distribution, σ is defined from the second- (μ_2) and first-order (μ_1) cumulant by $\sqrt{\mu_2}/\mu_1$. Thus, from Eq. (A4) we find that in a random text the value of

FIG. 2: Renormalised σ -index vs. Zipf's rank for each word in the Darwin's corpus (the first 4000 ranks). We have also plotted superposed the random version of the text (grey) and we have stressed the words corresponding to our prepared glossary (red online).

σ -index is given by

$$\sigma_{\text{ran}} = \sqrt{1 - p} . \quad (7)$$

Hence, we renormalise the index to eliminate this dependence on frequency defining

$$\sigma_{\text{nor}} = \frac{S}{\mu} \frac{1}{\sqrt{1 - p}} . \quad (8)$$

In Fig. 2 we plot the values of σ_{nor} for the first 4000 ranks in the Zipf's list of the Darwin's corpus. The random version of the corpus has also plotted in the same graph. The “cloud of points” corresponding to the random text is distributed around the unitary value of σ_{nor} , but the width of the “cloud” growth with rank. This fact is due to the frequency of occurrence decreases according the rank increases (Zipf's law), therefore the statistics gets worse. The words of our prepared version of the glossary are stressed in Fig. 2. From Fig. 2, it is appreciable that most of glossary words have high values of σ_{nor} .

B. Renormalised skewness

As σ , any cumulant contains partial information of the spatial distribution of words. Skewness is a parameter that describes the asymmetry of a distribution. Mathematically, the skewness is measured using the second- (μ_2) and third-order (μ_3) cumulant of the distribution according to $\kappa = \mu_3 / \mu_2^{3/2}$. Given that the distances between nearest neighbour tokens are positive defined, the corresponding distribution has positive skew, i.e., the upper tail is longer than the the lower tail.

From Eq. (A4) we find that in a random text the skewness of the distribution of distances between nearest neighbour tokens is given by

$$\kappa_{\text{ran}} = \frac{2 - p}{\sqrt{1 - p}} ; \quad (9)$$

Thus, the skewness also depends on the frequency of occurrence, p in the random case. We can define another renormalised quantity as we did with the σ -index, to eliminate this

FIG. 3: Renormalised κ -index vs. σ -index for *all* words in the Darwin’s corpus. We have also plotted superposed the random version of the text (grey) and we have stressed the words corresponding to our prepared glossary (red online).

dependence. In this manner, we write

$$\kappa_{\text{nor}} = \frac{\mu_3}{\mu_2^{3/2}} \frac{\sqrt{1-p}}{2-p} . \quad (10)$$

κ_{nor} can also be used for measuring relevance. However, the finite-size effects of the corpus are more pronounced for higher order cumulants. Instead, we use both cumulants σ_{nor} and κ_{nor} to construct a bi-dimensional graph for the corpus. Thus, in Fig. 3 we plot the the pairs $(\sigma_{\text{nor}}, \kappa_{\text{nor}})$ for *all* words in the Darwin’s corpus. In this graph, the “cloud of points” corresponding to the random text is distributed around the pair of values $(1, 1)$, while the region defined by $\sigma_{\text{nor}} > 2$ and $\kappa_{\text{nor}} > 2$ has almost none. The upper right corner of the graph concentrates almost all the points corresponding to the glossary. Figure 3 gives us immediately insight into the distribution of distances between nearest neighbour tokens, and provides us a graphical tool for determining keywords.

C. Renormalised Γ -index

As we did with the σ -index, we need to calculate Γ for a word type which appears in a random text with frequency p . For this task, we calculate the average of the random variable γ defined in Eq. (5) in a random text. From Eq. (A14) we obtain

$$\Gamma_{\text{ran}} = \frac{1}{2} h (h-1) (1-p)^h ((1-p) + (1-p)^{-1} - 2) , \quad (11)$$

where $h = \text{Int}[2/p]$. In this case, the dependence on p is ever more complicated than previous cases. This observation is absent from Ref. [19]. Zhou and Slater only calculate the value of Γ for the Poisson distribution: $\Gamma = 2e^{-2}$ (see Eq. (A17)), which is constant.

Now, as in the others cases, we define from Eqs. (6) and (11) a renormalised index by $\Gamma_{\text{nor}} = \Gamma/\Gamma_{\text{ran}}$. In Fig. 4 we plot the values of Γ_{nor} for the first 4000 ranks in the Zipf’s list of the Darwin’s corpus. The “cloud of points” corresponding to the random text is distributed around the unitary value but the width of the “cloud” growth with rank faster

FIG. 4: Renormalised Γ -index vs. Zipf's rank for each word in the Darwin's corpus (the first 4000 ranks). We have also plotted superposed the random version of the text (grey) and we have stressed the words corresponding to our prepared glossary (red online).

than in the case of σ_{nor} . The words corresponding to the glossary have systematically high values of Γ_{nor} .

IV. ENTROPY OF TOKEN DISTRIBUTIONS

Claude Shannon introduced the concept of *entropy of information* in 1948 [20]. Mapping a discrete information source on a set of possible events whose probabilities of occurrences are p_1, p_2, \dots, p_P , Shannon constructed a measure of information, choice and uncertainty, $S(p_1, p_2, \dots, p_P)$, requiring the following properties:

1. S should be continuous in the $\{p_i\}$.
2. For the iso-probability case, $p_i = 1/P$, S should be a monotonic increasing function of P .
3. If the set p_1, p_2, \dots, p_P is broken down into two subsets with probabilities $w_1 = p_1 + \dots p_k$ and $w_2 = p_{k+1} + \dots p_P$, then we must have the following composition law $S(p_1, \dots p_P) = S(w_1, w_2) + w_1 S(p_1/w_1, \dots p_k/w_1) + w_2 S(p_{k+1}/w_2, \dots p_P/w_2)$

The only S satisfying the three above assumptions is of the form

$$S = -K \sum_{i=1}^P p_i \log p_i, \quad (12)$$

where K is a positive constant.

A literary corpus can be divided in parts using natural partitions like parts, chapters, paragraphs or sentences. Thus, we consider the corpus as a composite of P parts. For the i -th part of the corpus we can reckon up the total number N_i of tokens in the part and the number $n_i(w)$ of occurrence of the word type w into this part. Then, the fraction $f_i(w) = n_i(w)/N_i$ ($i = 1, \dots, P$) is the frequency of occurrence of the word type w in the part i . Obviously, $\sum_{i=1}^P N_i = N$ is the total number of tokens in the corpus and $\sum_{i=1}^P n_i(w) = n(w)$

is the number of tokens corresponding to the word type w . Therefore, it is possible to define a probability measure over the partitions [14] as

$$p_i(w) = \frac{f_i(w)}{\sum_{j=1}^P f_j(w)} \quad (13)$$

The quantity $p_i(w)$ results more complex than the conditional probability $f_i(w)/(n(w)/N)$, of finding the word type w in the part i given that it is present in the corpus.

Following the Shannon's arguments, the information entropy associated with the discrete distribution $p_i(w)$ is

$$S(w) = -\frac{1}{\ln(P)} \sum_{i=1}^P p_i(w) \ln(p_i(w)) . \quad (14)$$

The value $1/\ln(P)$ for the constant K was selected to take the maximum value of S equal to one. Thus, $0 < S(w) < 1$. In this manner, when a type word is uniformly distributed ($p_i = 1/P$, for all i), Eq. (14) yields $S = 1$. Conversely, the another extreme case, $S = 0$, is when a word type appears in part j only, then we have $p_j = 1$ and $p_i = 0$ for $i \neq j$. Therefore, words with frequent grammatical use like function words (prepositions, adverbs, adjectives, conjunctions, and pronouns) will have high values of entropy, meanwhile keywords will have low values of entropy. Empirical evidence [14] shows a tendency of the entropy to increase with n . It implies that, on average, the more frequent word types are more uniformly used.

As we did with preceding indices, we need to calculate the average of the entropy of a mock word type that appears n times in a random corpus. From Eq. (B6) we obtain

$$(1 - S)_{\text{ran}} \approx \frac{P - 1}{2n \ln P} , \quad (15)$$

for $n \gg 1$ if all the parts of the random text have got the same number of tokens. Empirical evidence [14] shows that the agreement of Eq. (B6) with random shuffling of texts using natural partitions is very good, in spite of the limitation of the last assumption.

Montemurro and Zanette [14] proposed Eqs. (13) and (14) to study the distribution of words according to their linguistic role. For this task, they found that the suitable coordinates whereby words can be referred are $n(1 - S)$ and n . In the same way, we will use these ideas for detecting relevance of words. We cannot use directly the entropy as index because all tokens with only one occurrence have zero entropy. Thus, we define a normalised index freed from the dependence on frequency (n) in random texts by

$$E_{\text{nor}}(w) = n(w) (1 - S(w))_{\text{nor}} = n(w) \frac{2 \ln P}{P - 1} (1 - S(w)) . \quad (16)$$

FIG. 5: E_{nor} vs. number of occurrence (n) for each word in the Darwin’s corpus. We have also plotted superposed the random version of the corpus (grey) and we have stressed the words corresponding to our prepared glossary (red online).

Figure 5 shows the values of E_{nor} for all word types of Darwin’s corpus versus its number of occurrence, n , on a double logarithmic scale. The individual deviations from the bulk trend for each n are related to the particular usage nuances of words. For stress these deviations, we have used the 16 chapters of the corpus as natural partitions for our entropic analysis. In this way, we obtain a remarkable scattering of higher values of E_{nor} in the full range of number of occurrences. A same entropic analysis using the 842 paragraphs of the Darwin’s corpus as partitions generates a similar graph that stresses the bulk trend, but the fluctuations are completely smoothed. The “cloud of points” corresponding to the random version of the corpus is distributed around the unitary value and the corpus appears clearly more separated from the random text than previous cases. Additionally, the words corresponding to the glossary have systematically high values of the index E_{nor} . For reinforce our graphical findings, in the following section we perform a quantitative comparison among the indices σ_{nor} , Γ_{nor} , and E_{nor} based on the power of each index for discriminating the glossary from the bulk of words.

V. GLOSSARY AS BENCHMARK

The Origin of Species has got a Glossary of the principal scientific terms used in the book, prepared by Mr. W.S. Dallas, and an analytical index, both appended at the end of the book. Both sections, withdrawn from our prepared corpus, were written using 2418 word types. If we not consider the functions words, still remain 1679 word types (20% of the book’s lexicon). Then, we prepared by hand a version of the glossary selecting 283 word types (3.4% of the lexicon) with frequencies of occurrence greater than 9. The criterion for selection was rather more arbitrary, but we think that all selected words are pertinent to the book’s context. The graphs of glossary in Figs. 2–5 were drawn using this prepared set of 283 word types.

Now, we define *recall* [21] of an index of relevance as the fraction of glossary tokens

among the first top 283 ranked word types of the corpus. Recall is a useful benchmark for measuring the index’s performance. Thus, the performance of the index E_{nor} results 41%. In particular, the recalls of each index analysed in this work are

index	recall
E_{nor}	0.4169
σ_{nor}	0.4028
κ_{nor}	0.3781
Γ_{nor}	0.2544

In Table I we show the first top 50 word types of the prepared glossary ranked by the index E_{nor} . We also show the rank position of each word type by the others indices.

Another useful benchmark can be built looking for the last word type of our prepared glossary in the global ranking. It (FLOWERING) is ranked by the index E_{nor} in the position 2790. Remembering that the corpus has got 8294 word types, we obtain that the complete prepared glossary is allocated by E_{nor} in the first third part of the ranked lexicon.

A false positive is when the system identifies a keyword that really is not one. In Table II we show the first top 40 ranked (by E_{nor}) word types not included in our prepared glossary. We can immediately see that several terms are not necessarily false positives. With an asterisk (*), we have marked in the table those word types not previously selected in the prepared glossary, but that appearing in the main entries of the original index of the Darwin’s book. Indeed, several more word types like these could be included in our prepared glossary too. Moreover, the word type I is relevant for an autobiographical text like this book. ISLAND and SLAVES were not used neither in the book’s glossary nor in its index, however E_{nor} ranks it adequately as keywords. The word type F is also meaningful to the text. It appear in the proper nouns “Mr. F. Smith” and “Dr. F. Muller”, and in the collocations “F. sanguinea”, “F. rufescens”, “F. fusca”, “F. flava”, and “F. rufescens” which denote species. The observations in the last paragraphs induce us to consider that the performance of the index E_{nor} is better than the mentioned 41%.

Moreover the index E_{nor} requires less computational efforts that the others. Knowing the number of occurrences of a word type, the implementation of the algorithm for the variance requires of one accumulator plus a counter for reckoning the number of tokens between nearest neighbour occurrences. While, for the entropic index, we only need one counter (of

TABLE I: First top 50 word types of the prepared glossary ranked by the index E_{nor} .

Word type	E_{nor}^a	σ_{nor}^a	Γ_{nor}^a	Word type	E_{nor}^a	σ_{nor}^a	Γ_{nor}^a
HYBRIDS	1	2	13	SEA	33	65	309
STERILITY	3	1	7	SEEDS	35	64	279
SPECIES	5	447	1312	FERTILE	37	54	135
FORMS	6	185	667	ORGAN	39	14	218
VARIETIES	7	39	384	MOUNTAINS	40	120	94
INSTINCTS	8	3	19	GLACIAL	41	51	113
BREEDS	9	38	142	GARTNER	43	36	20
FERTILITY	10	8	33	HYBRID	44	46	59
FORMATIONS	11	20	78	CUCKOO	47	13	3
CROSSED	12	9	82	LAND	48	106	613
SELECTION	13	212	858	EGGS	50	109	215
ORGANS	14	61	433	STRUGGLE	51	829	571
NEST	16	22	18	BREED	52	332	367
INSTINCT	17	5	32	GEOLOGICAL	54	129	456
RUDIMENTARY	18	25	130	CROSS	62	125	205
FORMATION	19	144	341	HABITS	63	278	1260
BEEES	21	6	29	STRUCTURE	65	105	1451
PLANTS	22	113	776	INHABITANTS	67	95	556
CELLS	23	18	50	FLOWERS	68	35	250
POLLEN	24	12	74	ANTS	75	41	35
NATURAL	25	460	1288	RACES	78	566	542
GROUPS	26	79	393	OFFSPRING	81	400	884
CROSSES	27	60	81	SEXUAL	85	89	285
WATER	29	75	400	VARIABLE	87	138	467
STERILE	31	19	109	WILD	89	235	269

^aThe numerical values correspond to the ranking positions, not to the actual values of the indices.

TABLE II: First 40 false positives word types ranked by the index E_{nor} .

Word type	E_{nor}^a	n	Word type	E_{nor}^a	n
I	2	947	NORTHERN	60	41
ISLANDS	4	154 ^b	DESCENT	61	80 ^b
CHARACTERS	15	192 ^b	FRESH	64	50 ^b
GENERA	20	215 ^b	ITS	66	497
WAX	28	42	DIFFERENCES	69	168
ISLAND	30	69	CELL	70	30
DOMESTIC	32	131 ^b	EXTINCT	71	116 ^b
YOUNG	34	127	EUROPE	72	81 ^b
TEMPERATE	36	40	FERTILISED	73	34
SLAVES	38	34	DIAGRAM	74	40
NEW	42	278	SHALL	76	105
MY	45	99	WE	77	1320
INCREASE	46	82	DEVELOPED	79	146 ^b
INTERMEDIATE	49	164	BEDS	80	35
PERIOD	53	245 ^b	ADULT	82	46
MIVART	55	34 ^b	TWO	83	456
THROUGH	56	249	BETWEEN	84	367
HE	57	236	NUMBER	86	255
F	58	37	OCEANIC	88	42 ^b
PARTS	59	230 ^b	THEORY	90	131

^aThe numerical values correspond to the ranking positions, not to the actual values of the index.

^bWord types not previously selected in the glossary.

number of occurrences) for each partition per word type. The algorithm for Γ requires three accumulators and is much time-consuming because of for each occurrence of a word type we need to determine if it corresponds to a cluster point.

VI. CONCLUDING REMARKS

In summary, in this work we addressed the issue of statistical distribution of words in texts. Particularly, we have concentrated on the statistical methods for detecting keywords in literacy text. We reviewed two indices previously proposed [18, 19] for measuring relevance and we improved it by considering its values in random texts. Further, we proposed another index for keyword detection based on the entropy of information. Our proposal is very easy to implement and has got a performance as detector as good as the others indices.

The ideas of this work can be applied to any natural language with words clearly identified, but it is not required any previous knowledge about semantics or syntax.

Acknowledgements

Contributions to the B by Marcelo Montemurro are gratefully acknowledged. This work was partially supported by grant from “Secretaría de Ciencia y Tecnología de la Universidad Nacional de Córdoba” (Code: 05/B370).

APPENDIX A: THE GEOMETRICAL DISTRIBUTION

In this Appendix we briefly review the basic results of the geometrical distribution useful in this work. First, we consider an experiment with only two possible outcomes for each trial (binomial experiment). Repeated independent trials of the binomial experiment are called Bernoulli trials if their probabilities remain constant throughout the trials. We denote by p the probability of the “successful” outcome. Now, we are interested in the probability of success on the j -th trial after a given success. Given that the trials are independent, we immediately obtain the geometrical distribution

$$P(j) = (1 - p)^{j-1} p, \quad \text{for } j \geq 1. \quad (\text{A1})$$

1. Moments and cumulants

The characteristic function of a stochastic variable X is defined by $G(k) = \langle e^{kX} \rangle = \sum_{j \geq 1} P(j) \exp(kj)$. Thus, for the geometrical distribution we obtain

$$G(k) = \frac{p e^k}{1 - (1-p) e^k} . \quad (\text{A2})$$

This function is also the moment generating function

$$\langle X^n \rangle = \left. \frac{d^n G}{dk^n} \right|_{k=0} . \quad (\text{A3})$$

Therefore, the first three cumulants of the geometrical distribution are given by

$$\begin{aligned} \mu_1 &= \langle X \rangle = \frac{1}{p} , \\ \mu_2 &= \langle X^2 \rangle - \langle X \rangle^2 = \frac{1-p}{p^2} , \\ \mu_3 &= \langle X^3 \rangle - 3 \langle X^2 \rangle \langle X \rangle + 2 \langle X \rangle^3 = \frac{(2-p)(1-p)}{p^3} . \end{aligned} \quad (\text{A4})$$

2. Addition of two geometrical variables

If X_1 e X_2 are geometrical distributed independent random variables, the distribution of the addition $Y = X_1 + X_2$ is

$$P_Y(j) = \sum_{m_1+m_2=j} P(m_1, m_2) , \quad \text{for } j = 2, 3, \dots , \quad (\text{A5})$$

where the joint probability distribution, $P(m_1, m_2)$, is given by

$$P(m_1, m_2) = p^2 (1-p)^{m_1+m_2-2} , \quad \text{for } m_1 \geq 1, \text{ and } m_2 \geq 1 . \quad (\text{A6})$$

In this manner,

$$P_Y(j) = \sum_{m=1}^{j-1} P(m, j-m) = \sum_{m=1}^{j-1} p^2 (1-p)^{j-2} . \quad (\text{A7})$$

Therefore

$$P_Y(j) = (j-1) p^2 (1-p)^{j-2} , \quad \text{for } j = 2, 3, \dots . \quad (\text{A8})$$

Now, we are interested in the average of the random variable

$$\gamma = \begin{cases} 1 - \frac{Y}{2\mu} & Y < 2\mu \\ 0 & Y \geq 2\mu \end{cases} , \quad (\text{A9})$$

where Y is the addition of two independent geometrical distributed random variables with mean $\mu = 1/p$. By definition we have that

$$\langle \gamma \rangle = \sum_{j=2}^h \left(1 - \frac{j}{2\mu} \right) P_Y(n), \quad (\text{A10})$$

where $P_Y(n)$ is given by Ec. (A8) and $h = \text{Int}[2\mu]$. Defining $q = 1 - p$ and using the identity

$$\sum_{n=1}^N q^n = \frac{q - q^{N+1}}{1 - q} \quad (\text{A11})$$

we immediately obtain

$$\sum_{n=j}^h P_Y(j) = p^2 \frac{d}{dq} \sum_{k=2}^h q^{k-1} = 1 - h q^{h-1} + (h-1) q^h, \quad (\text{A12})$$

and

$$\begin{aligned} p \sum_{j=2}^h j P_Y(j) &= p^3 \frac{d^2}{dq^2} \sum_{k=2}^h q^k \\ &= 2 - h(h+1) q^{h-1} + 2(h+1)(h-1) q^h - h(h-1) q^{h+1}. \end{aligned} \quad (\text{A13})$$

Therefore

$$\langle \gamma \rangle = \frac{1}{2} h(h-1) q^h (q + q^{-1} - 2). \quad (\text{A14})$$

The Poisson distribution can be obtained from the geometrical distribution in the limit $p \rightarrow 0$. Expanding q^z into a Taylor series up to fourth order we obtain

$$q^{h+1} + q^{h-1} - 2q^h \approx p^2 + (1-h)p^3 + \frac{1}{2}(2-3h+h^2)p^4. \quad (\text{A15})$$

Given that for $p \rightarrow 0$ we have $h \gg 1$, the last equation can be recast as

$$q^{h+1} + q^{h-1} - 2q^h \approx p^2 \left(1 - hp + \frac{1}{2}(hp)^2 \right) \approx p^2 \exp(-hp). \quad (\text{A16})$$

Finally, using that $hp \approx 2$, we obtain the the average of the random variable γ for a Poisson distribution is

$$\langle \gamma \rangle = 2e^{-2} \quad (\text{A17})$$

APPENDIX B: ENTROPY OF A RANDOM TEXT

We consider a corpus of N tokens as a composite of P parts, with N_i tokens in the i -th part ($i = 1, 2, \dots, P$). In a random corpus, the probability that a word type w appears in

the part j is N_j/N . Thus, the probability that w appears n_1 times in part 1, n_2 times in part 2, and so on, is the multinomial distribution

$$p_w(n_1, n_2, \dots, n_P) = n! \prod_{j=1}^P \frac{1}{n_j!} \left(\frac{N_j}{N} \right)^{n_j}, \quad (\text{B1})$$

where $n = \sum_{j=1}^P n_j$ is the number of tokens corresponding to the word type w .

For reasons of simplicity, in this Appendix we consider the particular case in which all the parts have exactly the same number of tokens, i.e. $N_i = N/P$. Hence, the probability measure defined by Eq. (13) can be simply write as $p_i = n_i/n$ and the information entropy defined by Eq. (14) results

$$S = -\frac{1}{\ln P} \sum_{i=1}^P \frac{n_i}{n} \ln \left(\frac{n_i}{n} \right). \quad (\text{B2})$$

Now, we are interested in the average value of the entropy over the distribution given by Eq. (B1). We only need to compute the average of each term of Eq. (B2) using the marginal distributions, $p_w(n_i)$, obtained from Eq. (B1). All marginal distributions result binomials with mean n/P and variance $n/P(1 - 1/P)$. Thus, we obtain for the average entropy

$$\langle S \rangle = -\frac{P}{\ln P} \sum_{m=0}^n \frac{m}{n} \ln \left(\frac{m}{n} \right) \binom{n}{m} \frac{1}{P^m} \left(1 - \frac{1}{P} \right)^{n-m}. \quad (\text{B3})$$

For highly frequent word types, $n \gg 1$, we can approximate the binomial distribution by a Gaussian probability function ($G(x; \mu, \sigma)$) with mean $\mu = 1/P$ and variance $\sigma^2 = (1/n)(P - 1)/P^2$. Thus, Eq. (B3) can be recast as

$$\langle S \rangle \approx -\frac{P}{\ln P} \int_0^1 x \ln x G(x; \mu, \sigma) dx. \quad (\text{B4})$$

In the limit $n \gg 1$, $\sigma \rightarrow 0$ and the Gaussian probability function concentrates around its mean value μ . Using the expansion of the function $x \ln x$ around μ ,

$$x \ln x \approx \mu \ln \mu + (1 + \ln \mu)(x - \mu) + \frac{1}{2\mu}(x - \mu)^2, \quad (\text{B5})$$

in Eq. (B4) and remembering that $\int_{-\infty}^{\infty} (x - \mu)^2 G(x; \mu, \sigma) dx = \sigma^2$, we finally obtain for a random text [14]

$$\langle S \rangle \approx 1 - \frac{P - 1}{2n \ln P}. \quad (\text{B6})$$

-
- [1] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, (MIT Press, Cambridge, MA, 1999).
 - [2] H. P. Luhn, *The automatic creation of literature abstracts*, IBM J. Res. Devel. **2**, 159–165 (1958).
 - [3] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, (Addison-Wesley, Cambridge, MA, 1949).
 - [4] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, (McGraw-Hill, New York, 1983).
 - [5] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H. E. Stanley, *Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistic*, Phys. Rev. E **52**, 2939–2950 (1995).
 - [6] H. Stanley, S. Buldyrev, A. Goldberger, S. Havlin, C.-K. Peng and M. Simons, *Scaling features of noncoding DNA*, Physica A **273**, 1–18 (1999).
 - [7] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver and H. E. Stanley, *Analysis of symbolic sequences using the Jensen-Shannon divergence*, Phys. Rev. E **65**, 041905 (2002).
 - [8] W. Ebeling and T. Pöschel, *Entropy and long range correlations in literary English*, Europhys. Lett. **26**, 241–246 (1994).
 - [9] W. Ebeling, T. Pöschel and K.-F. Albrecht, *Entropy, transinformation and word distribution of information-carrying sequences*, Int. J. Bifurcation and Chaos **5**, 51–61 (1995) .
 - [10] M. Cassandro, P. Collet, A. Galves and C. Galves, *A statistical-physics approach to language acquisition and language change*, Physica A **263**, 427–437 (1999).
 - [11] A. Cohen, R. N. Mantegna and S. Havlin, *Numerical analysis of word frequencies in artificial and natural language texts*, Fractals **5**, 95–104 (1997).
 - [12] A. C.-C. Yang, C.-K. Peng, H.-W. Yien and A. L. Goldberger, *Information categorization approach to literary authorship disputes*, Physica A **329**, 473–483 (2003).
 - [13] R. F. i Cancho, *Decoding least effort and scaling in signal frequency distributions*, Physica A **345**, 275–284 (2005).
 - [14] M. A. Montemurro and D. H. Zanette, *Entropic analysis of the role of words in literary texts*,

- Adv. Complex Systems **5**, 7–17 (2002).
- [15] Digital text obtained from Project Gutenberg: <http://promo.net/pg>
 - [16] D. Beeferman, A. Berger and J. Lafferty, A model of lexical attraction and repulsion, in *Proceedings of the ACL-EACL Joint Conferences* (Madrid, Spain, 1997), pp. 373–380.
 - [17] T. Niesler and P. Woodland, Modelling word-pair relations in a category-based language model, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 2 (Munich, Germany, 1997), pp. 795–798.
 - [18] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz and A. M. Somoza, *Keyword detection in natural languages and DNA*, Europhys. Lett. **57**, 759–764 (2002).
 - [19] H. Zhou and G. W. Slater, *A metric to search for relevant words*, Physica A **329**, 309–327 (2003).
 - [20] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, (University of Illinois Press, Urbana, Illinois, 1949), reprinted with corrections from The Bell System Technical Journal **27**, pp. 379–423, 623–656, July, October, (1948).
 - [21] In informational retrieval, recall is defined as the proportion of the target items that a system recover [1].

This figure "figure1.png" is available in "png" format from:

<http://arxiv.org/ps/cs/0701028v1>

This figure "figure2.png" is available in "png" format from:

<http://arxiv.org/ps/cs/0701028v1>

This figure "figure3.png" is available in "png" format from:

<http://arxiv.org/ps/cs/0701028v1>

This figure "figure4.png" is available in "png" format from:

<http://arxiv.org/ps/cs/0701028v1>