Advances in Complex Systems, Vol. 5, No. 1 (2002) 7–17
© World Scientific Publishing Company

ENTROPIC ANALYSIS OF THE ROLE OF WORDS IN LITERARY TEXTS

MARCELO A. MONTEMURRO

Facultad de Matemática, Astronomía y Física, Universidad Nacional de Córdoba, Ciudad Universitaria, 5000 Córdoba, Argentina
mmontemu@famaf.unc.edu.ar

DAMIÁN H. ZANETTE

Consejo Nacional de Investigaciones Científicas y Técnicas, Centro Atómico Bariloche and Instituto Balseiro, 8400 Bariloche, Río Negro, Argentina
zanette@cab.cnea.gov.ar

Received 25 September 2001
Revised 29 October 2001
Accepted 29 October 2001

Beyond the local constraints imposed by grammar, words concatenated in long sequences carrying a complex message show statistical regularities that may reflect their linguistic role in the message. In this paper, we perform a systematic statistical analysis of the use of words in literary English corpora. We show that there is a quantitative relation between the role of content words in literary English and the Shannon information entropy defined over an appropriate probability distribution. Without assuming any previous knowledge about the syntactic structure of language, we are able to cluster certain groups of words according to their specific role in the text.

Keywords: Quantitative linguistics; language; word clustering; Shannon information entropy; Zipf's analysis.

1. Introduction

Language is probably the most complex function of our brain. Its evolutionary success has been attributed to the high degree of combinatorial power derived from its fundamental syntactic structure [3]. Syntactic rules act locally at the sentence level and do not necessarily account for higher levels of organization in large sequences of words conveying a coherent message [1]. In this respect, the situation is similar to that found in other natural sequences with non-trivial information content, such as the genetic code, where more than one structural level may be discerned [2, 8]. In the case of human language, complex hierarchies have been revealed at levels ranging from word form to sentence structure [1, 5, 6]. Moreover, it has been argued that long samples of continuous written or spoken language also possess a hierarchical macrostructure at levels beyond the sentence [9]. In a coarse grained splitting of this complex organization we can distinguish three basic structural levels in the analysis of long language records. The first corresponds to the absolute quantitative occurrence of words, that is, which words are used and how many times each. The second level of organization refers to the particular ways in which words can be linked into sentences according to the syntactic rules of language. Finally, at the highest level, grammatical sentences are combined in order to thread meaningful messages as part of a communication process. This assemblage of sentences into more complex structures is not strictly framed by a set of precise prescriptions and is more related to the particular nature of the message conveyed by the sequence. In this paper we shall focus on the statistical manifestations of this high level of organization in language. By means of an entropic measure of word distribution in literary corpora, we show that the statistical realization of words within a complex communicative structure reflects systematic patterns which can be used to cluster words according to their specific linguistic role.

2. Entropic Measure of Word Distribution

Zipf's analysis [10] represents the crudest statistical approach by which some quantitative information about the use of words in a corpus of written language can be obtained. Basically, it consists of counting the number of occurrences of each different word in the corpus, and then producing a list of these words sorted according to decreasing frequency. The rank-frequency distribution thus obtained presents robust quantitative regularities that have been tested over a large variety of natural languages. However, the frequency-ordered list alone carries little information on the particular role of words in the lexicon, as can be realized by noting that after shuffling the corpus the rank-frequency distribution remains intact. Naturally, the first ranks in the list belong to the commonest words in the language style of the source text, e.g. function words and some pronouns in literary English. After them, words related to the particular contents of the text start to appear. An illustration is given in Table 1, where we show some portions of the first ranks in Zipf's classification of the words from William Shakespeare's Hamlet.

It is therefore clear that in order to extract information about the specific role of words by statistical analysis, we must be able to gauge not only how often a word is used but also where it is used in the text. A statistical measure that fulfills the aforementioned requirement can be constructed from a suitable adaptation of the Shannon information entropy [7]. Let us think of a given text corpus as made up of the concatenation of P individual parts. The kind of partitions we are going to consider here are those that arise naturally at different scales as a consequence of the global structure of literary corpora. Some examples of these natural divisions are the individual books of an author's whole production, and the collection of chapters in a single book. Calling N_i the total number of words in part i, and n_i the number of occurrences of a given word in that part, the ratio $f_i = n_i/N_i$ gives

Table 1. Rank classification of words from Shakespeare's Hamlet.

Word	Rank	Number of occurrences
the	1	1087
and	2	968
to	3	760
of	4	669
I	5	633
a	6	567
you	7	558
...	...	...
Lord	25	225
he	26	224
be	27	223
what	28	219
King	29	201
him	30	197
...	...	...
Queen	42	120
our	43	120
if	44	117
or	45	115
shall	46	114
Hamlet	47	112
...	...	...

the frequency of appearance of the word in question in part i. For each word, it is possible to define a probability measure p_i over the partition as

$$p_i = \frac{f_i}{\sum_{j=1}^{P} f_j} \,. \tag{2.1}$$

The quantity p_i stands for the probability of finding the word in part i, given that it is present in the corpus. The Shannon information entropy associated with the discrete probability distribution p_i is

$$S = -\frac{1}{\ln P} \sum_{i=1}^{P} p_i \ln p_i \,. \tag{2.2}$$

Generally, the value of S is different for each word. As discussed below, the entropy of a given word provides a characterization of its distribution over the different partitions. Note that, independently of the specific values of p_i, we have $0 \leq S \leq 1$.

To gain insight into the kind of measure represented by S, two limiting cases are worth mentioning. If a given word is uniformly distributed over the P parts, $p_i = 1/P$ for all i and Eq. (2.2) yields $S = 1$. Conversely, if a word appears in part j only, we have $p_j = 1$ and $p_i = 0$ for $i \neq j$, so that $S = 0$. These examples represent extreme real cases in the distribution of words. To a first approximation one expects that certain words are evenly used throughout the text regardless of the specific contents of the different parts. Possible candidates are given by function words,

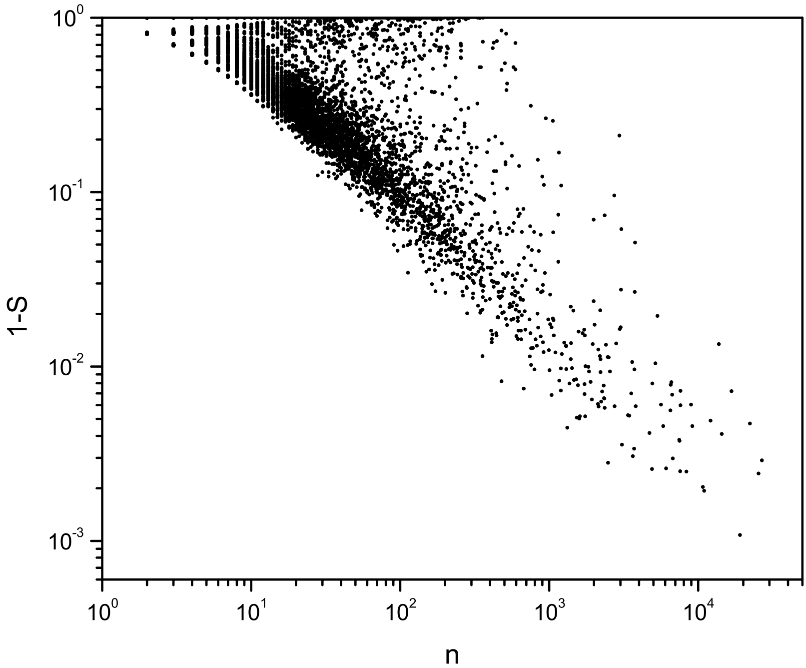

Fig. 1. Plot of $1 - S$ versus the number of occurrences n for each word of a corpus made up of 36 plays by William Shakespeare. The total number of words is 885,535 and the number of different words is 23,150.

such as articles and prepositions, whose use is only weakly affected by the specific character of the different parts in a homogeneous corpus. Other words, associated with more particular aspects of each part may fluctuate considerably in their use, thus having lower values of entropy. We show in the following that in just a few statistical quantities, such as frequency and entropy, there is relevant information about the role of certain word classes.

Figure 1 shows $1 - S$, with S calculated as in Eq. (2.2), versus the number of occurrences n, for each different word in a corpus made up of 36 plays by William Shakespeare. The total number of words in the set of plays adds up to 885,535 with a vocabulary of 23,150 different words. In this case, the natural division that we are considering is given by the individual plays. The structure of the graph calls for two different levels of analysis. First, its most evident feature, that is the tendency of the entropy to increase with n, represents a general trend of the data which should be explained as a consequence of basic statistical facts. In qualitative terms, it implies that, on average, the more frequent a word is the more uniformly is it used. Second, a somewhat deeper quest may be required in order to reveal whether the individual deviations from this general trend are related to the particular usage nuances of words, as imposed by their specific role in the text. Whereas most methods of word clustering according to predefined classes heavily rely on a certain amount of preprocessing, such as tagging words as members of particular grammatical categories [4], we shall address this point without any *a priori* linguistic knowledge, save the mere identification of words as the minimal structural units of language.

In order to clarify to which extent the features observed in Fig. 1 reflect basic statistical properties of the distribution of words over the different parts of the corpus, we performed a simple numerical experiment which consists of generating a *random version* of the 36 Shakespeare plays. This was done in the following steps. First, we considered a list of all the words used in the plays, each appearing exactly the number of times it was used in the real corpus. Second, we shuffled the list thus completely destroying the natural order of words. Third, we took the words one by one from the list and *wrote* a random version of each play containing the same number of words as its real counterpart. In Fig. 2, we compare the randomized version with the data of Fig. 1. It is evident that, on the one hand, the tendency of the entropy to grow with n is preserved. On the other, the large fluctuations in the value of S, as well as the presence of relatively infrequent words with very low entropy, are totally erased in the randomized version. On average, words have higher entropies in the random realization than in the actual corpus. Indeed, this is what one would expect for certain word classes such as proper nouns and, in

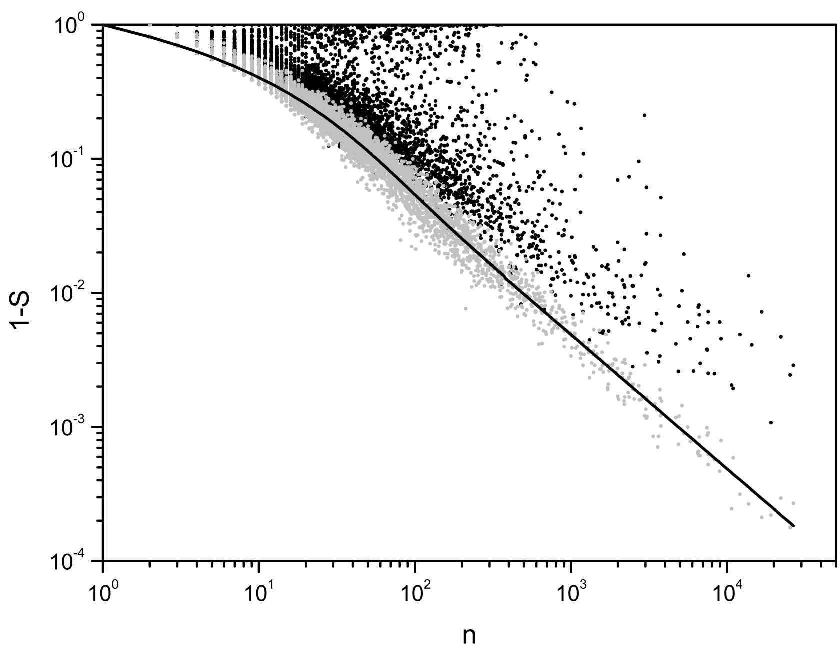

Fig. 2. Comparison between the data shown in Fig. 1 (black dots) and a randomized version of the Shakespeare corpus (grey dots). The curve represents the analytical approximation for the random corpus, Eq. (2.4).

general, for content words that allude to objects, situations or actions related to specific parts of the corpus. All the inhomogeneities that characterize the use of such words disappear in the random version, and consequently render higher values of the entropy.

Besides its value as a comparative benchmark, the random version of the corpus has the appeal of being analytically tractable, at least in a slightly modified form, as follows. Let us suppose that we have a corpus of N words consisting of P parts, with N_i words in part i ($i = 1, \ldots, P$). The probability that a word appears n_1 times in part 1, n_2 times in part 2, and so on, is

$$p(n_1, n_2, \ldots, n_P) = n! \prod_{j=1}^{P} \frac{1}{n_j!} \left(\frac{N_j}{N} \right)^{n_j}, \tag{2.3}$$

with $n = \sum_j n_j$. In the special case where all the parts have exactly the same number of words, i.e. $N_i = N/P$ for all i, the average value of the entropy resulting from the probability given by Eq. (2.3) can be written in terms of n only, as

$$\langle S(n) \rangle = -\frac{P}{\ln P} \sum_{m=0}^{n} \frac{m}{n} \ln \left(\frac{m}{n} \right) \binom{n}{m} \frac{1}{P^m} \left(1 - \frac{1}{P} \right)^{n-m}. \tag{2.4}$$

For highly frequent words, $n \gg 1$, Eq. (2.4) assumes a particularly simple form, namely

$$\langle S(n) \rangle \approx 1 - \frac{P-1}{2n \ln P}. \tag{2.5}$$

The curve in Fig. 2 stands for the function $1 - \langle S(n) \rangle$, with $\langle S(n) \rangle$ given by Eq. (2.4) with $P = 36$. First, we note that despite the fact that $\langle S(n) \rangle$ was calculated assuming that all the parts have the same number of words, its agreement with the random realization for the whole frequency range is very good. Moreover, it can be seen that after a short transient in the region of low n, $1 - \langle S(n) \rangle$ soon develops the asymptotic form given by Eq. (2.5) — a straight line of slope -1 in this log–log plot.

3. Distribution of Words According to Their Linguistic Role

We have so far explained the general trend in the behavior of the entropy with simple statistical considerations pertaining to the distribution of words over the different parts of the corpus. In the second part of our analysis we shall address the more interesting question of how the information contained in the entropy may reveal natural groupings of English words according to their particular role in the text. In order to accomplish this task it proves useful to define adequate *coordinates* whereby words can be associated to points in a suitable space. As a consequence of that, the classification of words according to their role should emerge naturally from the preference of certain words to occupy more or less definite regions of that space. As one of these coordinates we take a quantity reflecting the degree of use of a word in the text, namely, the number of occurrences n. The second coordinate is introduced to measure the deviation of the entropy of each word from the value predicted by the random-corpus model, as follows. We have seen that Eq. (2.5) accounts for the expected statistical decrease in the value of the entropy as a word becomes less frequent. This effect can be separated from the behavior of the words in the real texts, in order to reveal genuine information on the linguistic usage of words. We rewrite Eq. (2.5) as

$$(1 - \langle S \rangle)n \approx \frac{P-1}{2 \ln P}, \tag{3.1}$$

where the right-hand side is independent of n. Therefore, in a graph of $(1 - S)n$ versus n, the words whose actual distribution agrees with the random-corpus model should approximately fall along a horizontal line. All appreciable departures from this line should be expected to bear some relation to the non-random character of the usage of words, and therefore may reflect actual linguistic information. By means of relation (3.1) we are therefore able to filter out all the trivial parts of the statistical behavior. Figure 3 shows actual data for the Shakespeare corpus in a plot of $(1 - S)n$ versus n. The horizontal line stands for the value given in the right-hand side of Eq. (3.1) for $P = 36$.

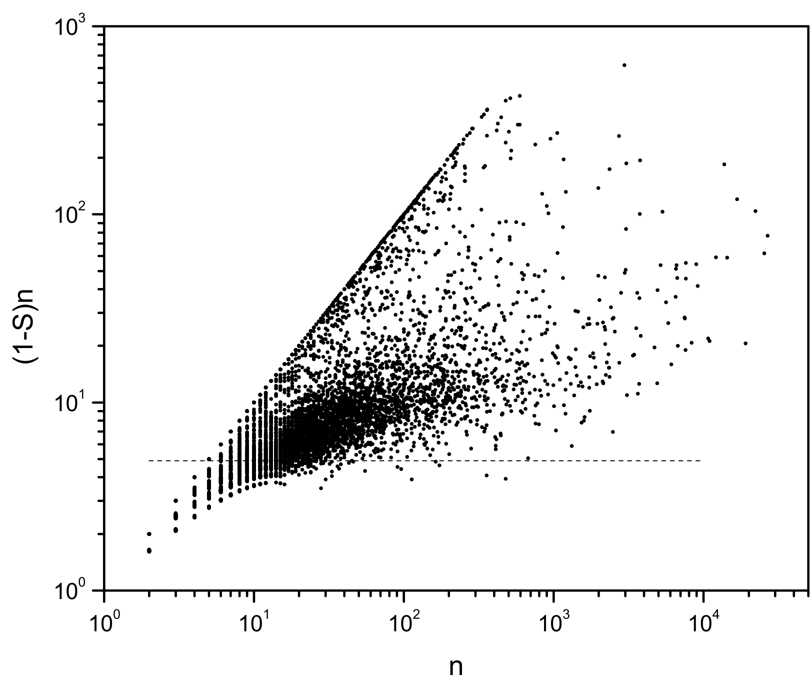

Fig. 3. Plot of $(1 - S)n$ versus n for all the different words in the Shakespeare corpus. The horizontal line shows the expected value of $(1 - S)n$ for frequent, uniformly distributed words, as given by Eq. (3.1).

In order to reveal whether the words show some sort of systematic distribution over the plane according to their linguistic role, we proceeded to classify by hand the first 2,000 words into different sets. The classification stops there due to the fact that words close to that rank occur in the whole corpus a number of times similar to the number of parts in the corpus division, $n \approx P$, thus representing a limit beyond which statistical fluctuations start to dominate. The six categories we arranged into groups were the following: (a) proper nouns, (b) pronouns, (c) nouns referring to humans, such as *soldier* and *brother*, (d) nouns referring to nobility status, such as *King* and *Duke*, which have a relevant place in Shakespeare's plays, (e) common nouns (not referring to humans or to nobility status) and adjectives, and finally (f) verbs and adverbs. In the case of ambiguity about the inclusion of a word into a certain class we simply left it out and did not classify it, hence the total number of classified words was finally around 1,400.

The results of this classification can be seen in Fig. 4, and in fact reveal a marked clustering of words over definite regions of the two-dimensional space spanned by $(1 - S)n$ and n. The sharpest distribution, shown in Fig. 4(a), corresponds to proper nouns. These words occupy a dense and elongated region which is limited from above by the straight line representing the identity function $(1 - S)n = n$. Naturally, proper nouns are expected to define a class of words strongly related to particular parts of the corpus. In consequence, their entropies tend to be very low on average, if not strictly zero as in the case of many proper nouns appearing in just one of Shakespeare's plays. Therefore, in a graph of $(1 - S)n$ versus n, words having values of entropy close to zero have $(1-S)n \approx n$ and fall close to the identity function.

The distribution of other word classes is less obvious. Verbs and adverbs (Fig. 4(f)) are closest to the random distribution, covering a wide range of ranks. On average, common nouns and adjectives (Fig. 4(e)) are farther from the random distribution and, at the same time, are less frequent. Nouns referring to humans (Fig. 4(c)) cover approximately the same frequencies as common nouns, but their distribution is typically more heterogeneous. The entropy of some words in this class is, in fact, quite close to zero. The three most frequent nouns in the Shakespeare corpus are *Lord*, *King* and *Sir*. All of them belong to the class of nouns referring to nobility status (Fig. 4(d)), which spans a large interval of frequencies and has systematically low entropies. The specificity of nobility titles with respect to the different parts of the corpus can be explained with essentially the same arguments as for proper nouns. Considerably more surprising is the case of pronouns (Fig. 4(b)) which, as expected, are highly frequent, but whose entropies reveal a markedly nonuniform distribution over the corpus. The origin of these heterogeneities in the distribution of some word classes is not at all clear, and deserves further investigation in the frame of linguistics. In Fig. 5 we have drawn together the zones occupied by all the classes to make more clear their relative differences in frequency and homogeneity.

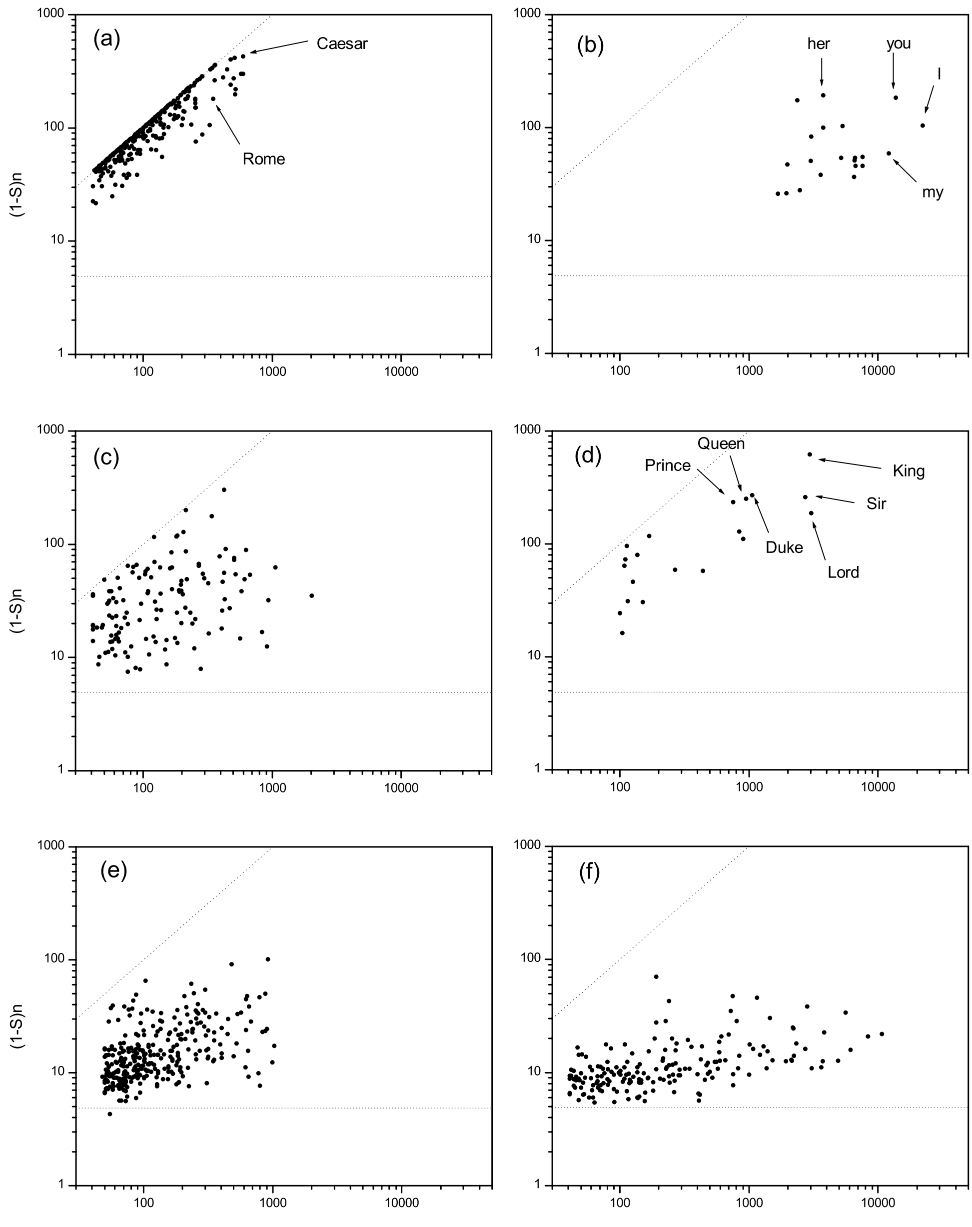

Fig. 4. Plot of $(1 - S)n$ versus n for six relevant word classes: (a) proper nouns, (b) pronouns, (c) nouns referring to humans, (d) nouns referring to nobility status, (e) common nouns (not referring to humans or to nobility status) and adjectives, and (f) verbs and adverbs. In each plot, the horizontal dotted line stands for the asymptotic value of $(1-S)n$ for the random-corpus model, Eq. (2.5). Words close to this line are homogeneously distributed over the corpus. The oblique dotted line corresponds to $S = 0$. Proximity to this line indicates extreme inhomogeneity in the distribution.

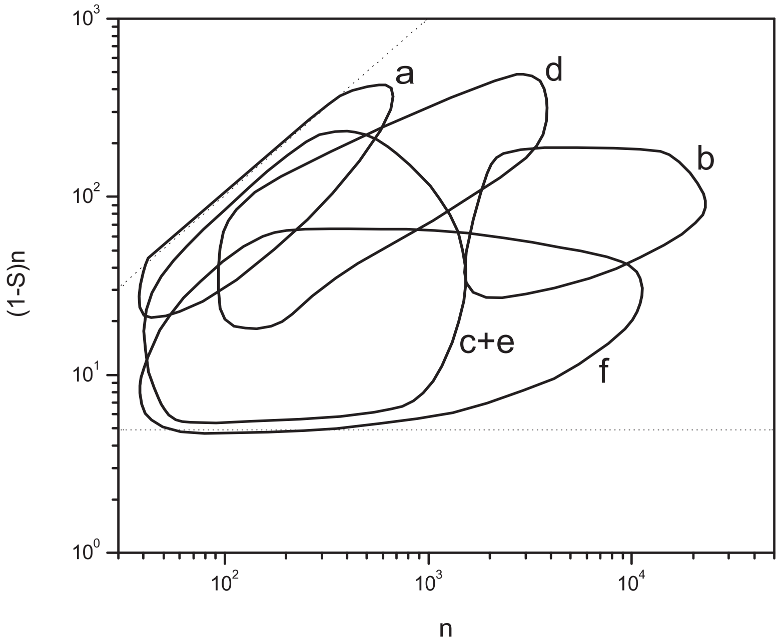

Fig. 5. Schematic combined representation of the zones occupied by the word classes of Fig. 4. Labels and scales are as in Fig. 4.

4. Conclusion

We have performed the same statistical analysis over other literary corpora, obtaining totally consistent results. The same organization of words was observed in the works of Charles Dickens and Robert Louis Stevenson. In particular, nouns and adjectives tend to be more heterogeneously distributed than verbs and adverbs. Nouns referring to humans have systematically lower entropies. Pronouns, in turn, exhibit an unexpectedly heterogeneous distribution for their high frequencies. Though we are not able to give a full explanation for such heterogeneities on linguistic grounds, we expect that the disclosure of these features will incite further analysis by specialists.

In summary, in this work we have concentrated on the statistical analysis of language at a high level of its structural hierarchy, beyond the local rules defined by sentence grammar. We started off by introducing an adequate measure of the entropy of words in a text corpus made up of a number of individual parts. With respect to Zipf's analysis, which focuses on the frequency distribution of words, the study of entropy provides a second degree of freedom that resolves the statistical behavior of words in connection with their linguistic role. By means of our random-corpus model we were able to extract the nontrivial part of the distribution of words. This procedure reveals statistical regularities in the distribution, that can be used to cluster words according to their role in the corpus without assuming any *a priori* linguistic knowledge. Ultimately, such regularities should stand as a manifestation of long-range linguistic structures inherent to the communication process. We believe that a thorough explanation of the origin of these global structures in language may eventually contribute to the understanding of the psycholinguistic basis for the modelling of reality by the brain.

Acknowledgments

Critical reading of the text by Susanna Manrubia is gratefully acknowledged. Marcelo A. Montemurro receives financial support from the Secretary of Science and Technology of the Universidad Nacional de Córdoba, Argentina.

References

[1] Aknajian, A., Demers, R. A., Farmer, A. K. and Harnish, R. M., *Linguistics. An Introduction to Language and Communication* (MIT Press, Cambridge, 1992).

[2] Daune, M., *Molecular Biophysics. Structures in Motion* (Oxford University Press, Oxford, 1999).

[3] Nowak, M. A., Plotkin, J. B. and Jansen, V. A. A., *Nature* **404**, 495 (2000).

[4] Pereira, F., Tishby, N. and Lee, L., Distributional clustering of English words, in *30th Annual Meeting of the Association for Computational Linguistics* (Ohio State University, Association for Computational Linguistics, Columbus, Ohio, 1993), pp. 183–190.

[5] Pinker, S., *Words and Rules* (Harper Collins, New York, 2000).

[6] Pinker, S., *The Language Instinct* (Harper Collins, New York, 2000).

[7] Shannon, C. E., *The Bell System Technical Journal* **27**, 379 (1948).

[8] Smith, J. M. and Szathmáry, E., *The Origins of Life* (Oxford University Press, Oxford, 2000).

[9] Van Dijk, T. A., Semantic macro-structures and knowledge frames in discourse comprehension, in *Text and Context*, Just and Carpenter, eds. (Longmans, London, 1977).

[10] Zipf, G. K., *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Reading, 1949).