Eur. Phys. J. B **49**, 523–529 (2006)
DOI: 10.1140/epjb/e2006-00091-3

THE EUROPEAN
PHYSICAL JOURNAL B

The network of concepts in written texts

S.M.G. Caldeira[1], T.C. Petit Lobão[1], R.F.S. Andrade[1,a], A. Neme[2], and J.G.V. Miranda[1]

[1] Instituto de Física, Universidade Federal da Bahia, Campus Universitário da Federação, 40210-340, Salvador, BA, Brazil
Instituto de Matemática, Universidade Federal da Bahia, Campus Universitário da Federação, 40210-340, Salvador, BA, Brazil

[2] Institut Gaspard-Monge, Université de Marne-la-Vallée, 77454 Marne-la-Vallée Cedex 2, France

Received 14 November 2005
Published online 31 March 2006 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2006

Abstract. Complex network theory is used to investigate the structure of meaningful concepts in written texts of individual authors. Networks have been constructed after a two phase filtering, where words with less meaning contents are eliminated and all remaining words are set to their canonical form, without any number, gender or time flexion. Each sentence in the text is added to the network as a clique. A large number of written texts have been scrutinised, and it is found that texts have small-world as well as scale-free structures. The growth process of these networks has also been investigated, and a universal evolution of network quantifiers have been found among the set of texts written by distinct authors. Further analyses, based on shuffling procedures taken either on the texts or on the constructed networks, provide hints on the role played by the word frequency and sentence length distributions to the network structure.

PACS. 89.75.Fb Structures and organization in complex systems – 89.75.Hc Networks and genealogical trees – 02.10.Ox Combinatorics; graph theory

1 Introduction

Concepts of complex networks have proven to be powerful tools in the analysis of complex systems [1–5]. They have been applied to modelling purposes as well as to search for properties that naturally emerge in actual systems due to their large-scale structure. Unlike random graphs, complex networks reveals ordering principles related to their topological structure. This way, if complex systems are mapped onto networks, it is possible to use their conceptual framework to identify and even to explain features that seem to have universal character. Several complex networks have been proposed in the scientific literature associated with real systems: the biological food web [6], technological communication networks as the Internet, information networks as the World Wide Web [7], social networks defined by friendship relations among individuals, etc. [8].

Word networks have been used to address complex aspects of human language. In such studies, words are connected according either to semantic associations [11–13] or even by nearness in the text [14–17], i.e., based on what is commonly called word window with a fixed number of words. Those works intend to establish the structure of a given language as a whole. Because of this, they deal with a huge amount of texts, independently of their authors, what is called *corpora*.

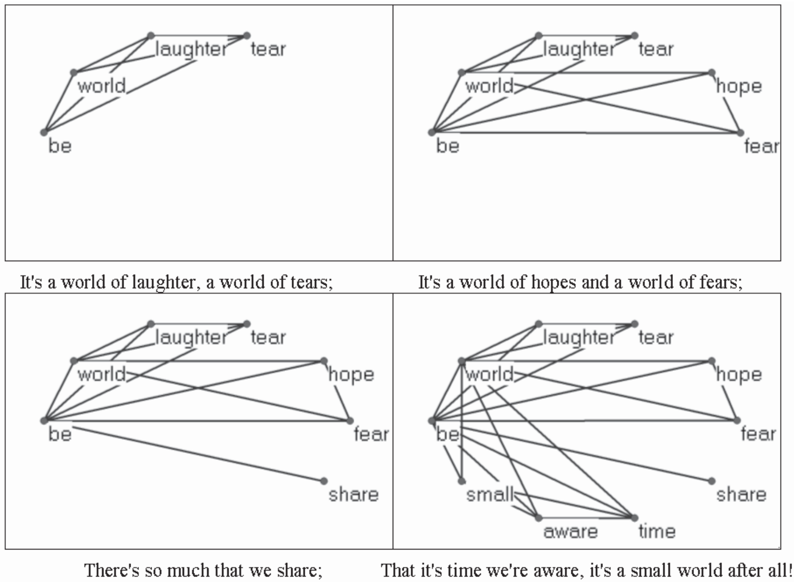
As a product of brain activity, language and language networks can be brought in connection to neuronal networks revealed by direct physiological measurements [18]: functional magnetic resonance of human brains established networks whose vertices are regions of the cerebral cortex activated by external stimuli according to a temporal correlation of activity.

In this work, we investigate the relations between the concepts in individual written texts, by using them as starting point to construct significant networks. Projecting both the concepts present in the text, as well as the way they are related among them, onto a network gives the opportunity to use the tools and concepts developed within the network framework to characterise, in a quantitative way, how the concepts in a written text appear, how ordered and connected they are, how close to each other they are within the text, and so on.

2 Text network construction

An undirected network is defined by a set of elements, called vertices or nodes, represented graphically by points, some of which are joined together by an edge, represented by a line between these points. The topological structure

[a] e-mail: randrade@ufba.br



It's a world of laughter, a world of tears;

It's a world of hopes and a world of fears;

There's so much that we share;

That it's time we're aware, it's a small world after all!

Fig. 1. The filtering treatment and the growth process of the network for a small poem.

of many networks displays a large-scale organization, with some typical properties like: highly sparse connectivity, small minimal paths between vertices and high local clustering.

To analyse a network, a number of indices (or quantifiers) have been proposed, which allow for a quantification of many of the properties quoted above. In this work we shall use the most important of them: number of vertices N, number of edges M, average connectivity $\langle k \rangle$, average minimal path length ℓ, diameter L, degree distribution $p(k)$, and the Watts and Strogatz clustering coefficient C (for a thorough discussion of these indices see, e.g., [5]). If the node degree distributions follow power laws, $p(k) \sim k^{-\gamma}$, the exponent γ becomes an important index for the network characterisation. Two important network scenarios associated, respectively, to the small-world [9] and scale-free structures [10], have been identified in several complex networks, including those related to the subject of this work, namely in the quoted *corpora* and of brain activity.

To map the texts onto networks that reflect the structure of meaning concepts within it, we have preserved only the words with an intrinsic meaning, eliminating words which have a merely grammatical function, as to arrange the syntactical structure of sentences in the text (articles, pronouns, prepositions, conjunctions, abbreviations, and interjections). Afterwards, we have reduced the remaining words to their canonical forms, i.e., we have disregarded all inflections as plural forms, gender variations, and verb forms. Such procedures are common in studies on language [19]. In order to perform a computer implementation, we have used some routines, dictionaries, and grammatical rules from UNITEX package [20]. Unknown words to the program were preserved in their original forms. After this filtering treatment, we have constructed a network for each individual text. The construction is based on the concept of sentence, which is considered as the smallest significant unit of the discourse. To be more precise, in this work sentence is defined as a string of words limited by two full stops, indicated by any of these graphic signs: period, colon, question mark, exclamation point, ellipses. In the network, each distinct word corresponds to a single vertex, and any two words are connected if they are concomitantly present in one (or more) same sentence. Sentences are incorporated into the network as a complete subgraph, that is, a clique of mutually connected words. Sentences with common words are connected by the shared words. The method we propose offers a natural way to analyse the growth process of the network evolution. We have investigated both the network behavior in this step-by-step evolving process, what we call dynamical analysis, as well as the behavior of the entire network in its final form, corresponding to the entire text, called the static analysis. Both methods reveal important properties and they shall be discussed further on. In Figure 1, we show an example of a text filter process together with the evolving process of network construction for a well known jingle.

In order to guarantee an uniform and comprehensive sample, we have chosen a collection of 312 texts [21], which can be cast into different classes as: genre (59% technical, 41% literary), language (53% in Portuguese, 47% in English), gender (72% male authors, 28% female authors). Finally we have also classified the texts according to their size (55% with less than 1000 sentences, 45% with more than 1 000). The smallest text has 169 and the largest one, 276 425 words; in the average, the texts have 32 691 words.

3 Results

3.1 Text analysis

The texts were individually analysed, based on the evaluation of the indices quoted in the previous section: N,

Table 1. Average values for the indices. Filtered texts were used as standard subjects, and shuffling operations were carried out on networks generated by them. Original means text without filtering, used just for the purpose of comparison.

Text	L	ℓ	C
Filtered	5 ± 1	2.3 ± 0.2	0.77 ± 0.05
SA	4 ± 1	2.2 ± 0.2	0.77 ± 0.04
SB	4 ± 1	2.4 ± 0.3	0.74 ± 0.05
SC	4 ± 1	2.2 ± 0.3	0.40 ± 0.20
SD	3 ± 1	2.2 ± 0.3	0.05 ± 0.06
Original	4 ± 1	2.0 ± 0.1	0.82 ± 0.03

M, $\langle k \rangle$, ℓ, C, L, and degree distribution $p(k)$, which we have found to obey a power law $p(k) \sim k^{-\gamma}$. We have also followed how the number and size of independent clusters varies in the growth process. Computed average values of such indices based on all investigated texts are shown in Table 1. The analysed networks are very sparse, with average connection index $2M/[N(N-1)] < 0.1$ for the networks with $N \geq 100$.

We have verified that the degree distributions, for all analysed texts, exhibit an early maximum around $k = 10$, which is followed by crossover to a long tail, approximating a power law distribution with average exponent $\langle \gamma \rangle = 1.6 \pm 0.2$. We have found no evidence of correlation between γ and the length of the analysed text. These features are illustrated in Figure 2, where we show $p(k)$ versus k for a very large text and the smallest one we analysed.

In Figure 3a, we indicate by squares the value of clustering coefficient for the 312 analysed texts as function of the text size. The obtained values are always large, lying in the interval $[0.68, 0.9]$. We also see that texts with smaller number of vertices have, in general, higher C values in comparison to those with larger number of vertices. The dynamical analysis of the evolution of C reveals interesting results that may help understand this behaviour. For the purpose of comparison, we superimpose in Figure 3a a sequence of values of C, evaluated as the construction of the network evolves, for two different very large texts. We clearly see that C goes through a maximum, and then decreases to a text dependent value $C_{0,i}$ which, for the two examples, are ~0.75. So, the dynamical dependence of C on the text length for a single text follows the average trend observed for an set of texts.

A similar behaviour for C is exhibited when we plot its value as function of the number of sentences s in 13 different texts, as shown in Figure 3b. There we draw $C - C_{0,i}$ as function of the number s of sentences in the text in linear logarithmic scale, what indicates that as $C \approx C_{0,i}$ $\exp(-s/\sigma)$.

The investigation of the cluster structure of the network during its growth has shown that texts evolve mostly in the form of a very large single cluster, which is formed since the very beginning of the network evolution. New words are likely to adhere to it rather than starting other significant clusters, what leads to the presence of a single giant cluster for the entire text. Another aspect we have analysed was the evolution of the indices in the network growth process with respect to the same text in different languages. As an example we compare, in Figure 4, the evolution of C and ℓ for the network associated to James Joyce's Ulysses, in the original English version and in its translation to Portuguese.

The results we have obtained to this point indicate that the networks we analysed have highly sparse connectivity, small D and ℓ, but high C, what constitute evidences of a small-word network scenario. Besides, since $p(k)$ decay according to power laws, we conclude they also behave as scale-free networks. It is important to emphasize that the filtering treatment does not modify the network general behavior. In order to verify this fact, we have also performed the same measurements for the networks obtained from the original texts, without any kind of treatment. We have verified that, notwithstanding to the fact that we obtain different values for the same indices (see Tab. 1), the very frequent presence of articles, prepositions and other words does not alter the small-world and scale-free character of the text networks! The small-world behavior is characterized by a great compactness and by the presence of some vertices with high local clustering. Such aspect in these networks means that they have words that are often used along the text. That fact explains the small values of D and ℓ. The scale-free behavior results from the presence of nodes with high connectivity degree in a much greater amount in comparison to that of a random graph. This aspect is obviously expected in a written text, due to the presence of an organization principle which controls the construction of the text. The most connected words are associated with recurrent concepts around which the text is constructed.

3.2 Shuffling procedures

There are several agents that contribute to determine the measured indices in the analysed networks. For example, since each sentence is added to the network as a complete subgraph, their lengths may interfere in the clustering coefficient. In the same way, the own structure of the sentences and the frequency of the words along the whole text affects C and other indices as well. In order to evaluate the role of these main agents in the determination of the indices, we have re-analysed the filtered texts after submitting them to four shuffling processes, which are so characterized: (*SA*) The beginning and the end of sentences are kept fixed, while the position of the words in the whole text is randomly changed. It breaks the structure of concepts in the sentences, but sentence lengths and word frequencies are kept unchanged. (*SB*) The original sequence of the words along the text remains unchanged, but all sentences are forced to have the same average length, obtained from the unperturbed text. (*SC*) The beginning and the end of sentences are kept fixed, and words are randomly chosen from the same vocabulary of the text. All the words have the same choice probability.

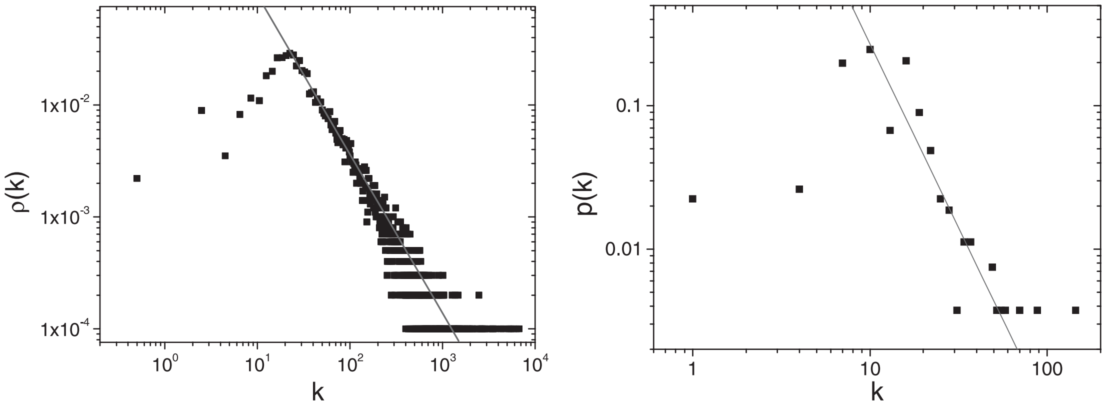
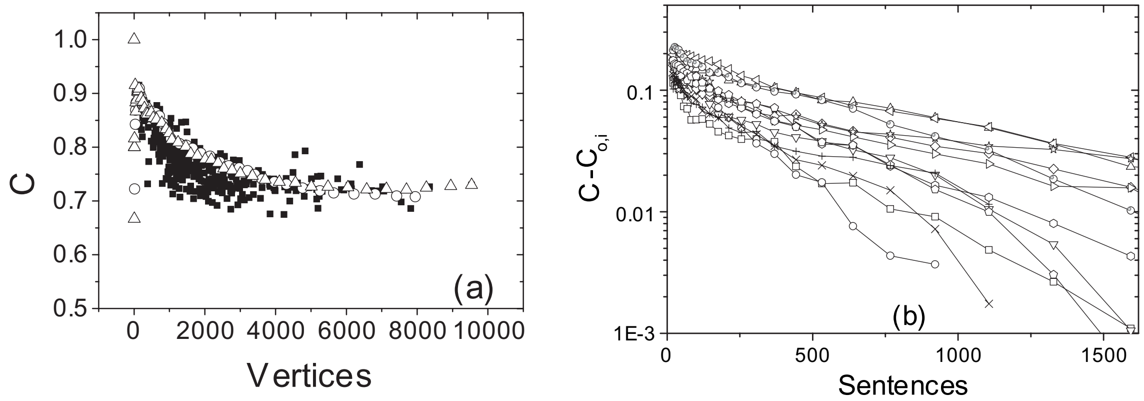


Fig. 2. Degree distributions $p(k)$ for a very large and the smallest text.



(a) (b)

Fig. 3. (a) Behavior of C as function of the number of vertices for 312 texts compared with the evolution of two large texts: Dan Brown's The da Vinci Code (hollow circles) and Tolstoi's Anna Karenina (hollow triangles). (b) $C - C_{0,i}$ as function of the number of sentences, where i labels the 13 texts.

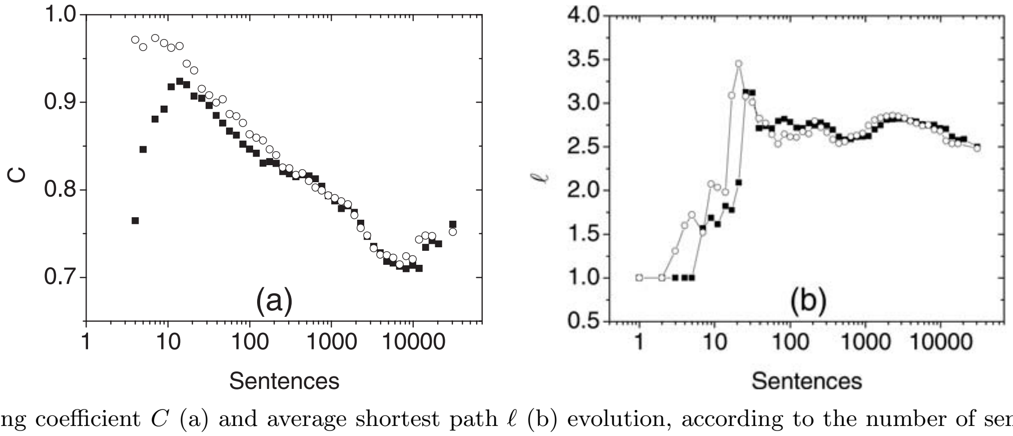


Fig. 4. Clustering coefficient C (a) and average shortest path ℓ (b) evolution, according to the number of sentences in James Joyce's Ulysses, in the original English version (full squares) and in a Portuguese translation (hollow circles).

(*SD*) Erdös-Renyi network with the same number of nodes and links as in the filtered text networks.

A summary of results is also included in Table 1. They show that the networks are affected in different ways but, as expected, for all but the *SD* procedure, they are very far from random networks. The indices for *SA* and *SB* are not significantly altered. In the first case it shows that, without changing the structural aspects of sentence sizes and word frequencies, the network is not essentially affected. In the second one, breaking sentence sizes but keeping words in their original places is not sufficient to alter, in a meaningful way, the network structure. This behavior is shown in Figure 5, where 4 cumulative degree distribution

$$P(k) = \frac{1}{k}\int_k^\infty p(k')dk', \tag{1}$$

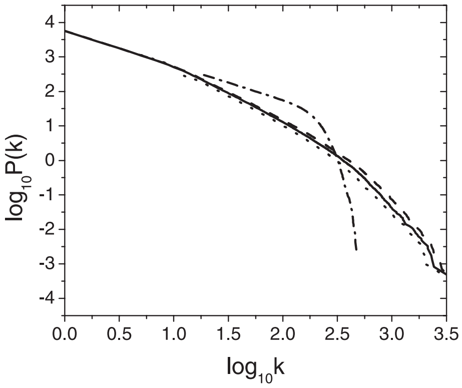


Fig. 5. Cumulative degree distributions $P(k)$ obtained with one single text (Bio-informatics for Geneticists, by Barnes and Gray) for the filtered version (solid), and the shuffle procedures *SA* (dashed), *SB* (dotted), and *SC* (dash-dotted).

for a given text and three shuffled versions of it are drawn. We realize that the scale-free character, expressed by a linear decay in the k interval [10, 300] for $P(k)$ remains almost unchanged. These results remind us of Zipf's universal result about distribution frequency of words in texts since, when Zipf's law is not followed, e.g., *SC* and *SD*, the network structure is broken. Indeed, not only the power law for $P(k)$ is lost, but also the indices in Table 1 have been altered. Nevertheless we emphasise the fact that, despite the change in word frequency caused by *SC* sets up a deep modification in the text, its value for C is much higher than that one for *SD*. This indicates that the unchanged distribution of sequence sizes still keeps the *SC* network far away from the Erdös-Renyi scenario.

3.3 Addition of new words

To conclude this section let us discuss the dependence of the rate at which new words are added to the text as it is being written. As expected, the measured probability $P(w; L)$ that a new word w is added decreases while the length L of the text increases. We have estimated this probability by simply counting the average number of new words in the sentences, as shown in Figure 6. There, we draw $P(w; L)$ for two texts where this general feature is explicitly shown. The first three entries of the series stay for the the first sentences of any text, and correspond to title, name of the author and title of the first chapter. After a short transient phase due to the first sentences, which actually is devoid of statistical significance, we always find a decreasing behaviour that can be described by a power law

$$P(w; L) \sim L^{-\zeta}, \qquad (2)$$

where $\zeta \in (0.25, 0.4)$ for the texts we analysed.

4 A model for network growth

The results of the dynamical analysis call our attention to general aspects of network growth. Barabási and

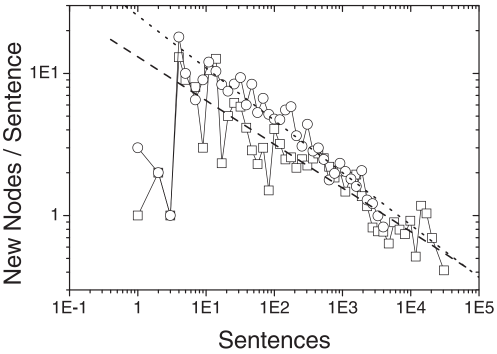


Fig. 6. Evolution for the average number of new words for sentence in two texts (Ulysses (by Joyce) — squares; and Genetic Nature/Culture: Anthropology and Science Beyond the Two-Culture Divide (by Goodman et al.) — circles.)

Albert [10] observed that the scale-free behavior may be explained by a special kind of growth process, known as preferential attachment. They proposed a model that captures such behavior: vertices are added to the network, systematically, by connecting them to some just existing vertices, which are selected in accordance with a probability distribution that depends on their degrees. However, the values obtained for C are much lower than those in small-world networks. On the other hand, the Watts and Strogatz small-world, obtained by randomly rewiring a regular network, does not reproduce the scale-free feature [9]. Our results suggest that the growth process we employ here is essential to capture both quoted behaviors. This process is characterised by the addition of a new sentence, i.e., a complete subgraph or clique in each step. Due to the frequency distribution of words along the text, the attachment of these complete subgraphs is still preferential, but the new vertices are highly clustered, what contributes to the coexistence of both small-world and scale-free network scenario.

A simple model was set up to help us understand the general features of the results obtained from the analysed texts. It has some common features with Barabasi's model of preferential attachment. However, an essential difference refers to the fact that the network grows by attachment of new cliques, not individual vertices. New added cliques, which have a characteristic mean length, are intended to describe the inclusion of a new full sentence. Another important difference refers to the fact that new cliques are composed of new and old vertices (words). This is necessary in order to keep pace with corresponding feature observed in the growth of the text networks. To account for the relative presence of new to old words in the added sentences, as discussed in the previous section, we let the average ratio of new to old vertices in the cliques decreases as the construction of the network proceeds, according to the law expressed in equation (2).

Our model depends on three parameters: the total number Q of cliques (sentences), the largest number M of words in a sentence, and the exponent that describes the inclusion of new vertices ζ. The sizes of new clique added to the network are randomly chosen in the interval

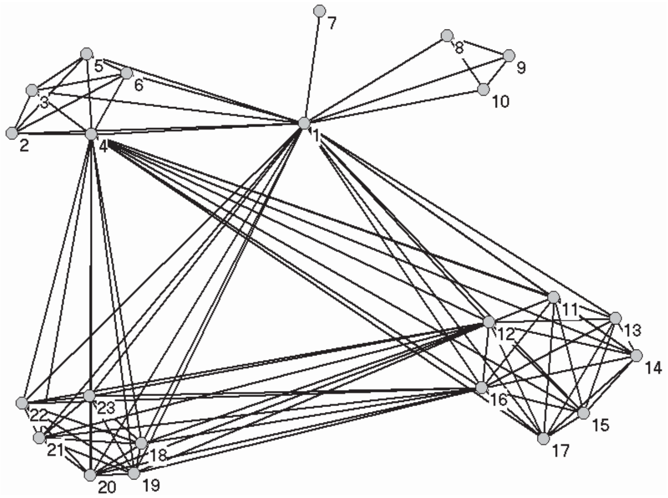


Fig. 7. An example of the model for six sentences with maximum size of ten words and $\zeta = 1/4$.

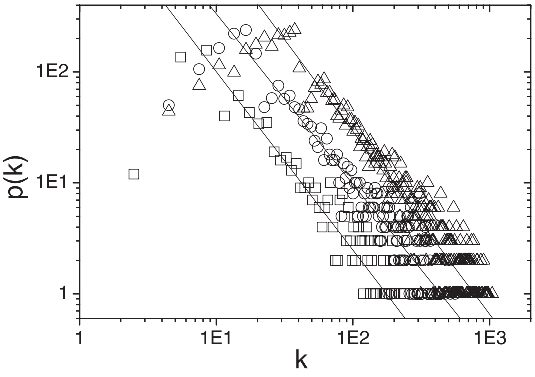


Fig. 8. Distribution degree for tree simulated networks with 1000 sentences, $\zeta = 1/4$ and with different values of maximum sentences sizes. The lines represent linear regression fitting in the straight line region.

$[1, M]$. The number of new and old vertices in each added clique is chosen according to (2), while the choice of the old vertices included is made on the basis of the preferential attachment mechanism, i.e., hubs are more likely to be included in the cliques than nodes with only a few connections. This procedure simulates the growth of our text based networks. In Figure 7 we illustrate the growth of a network for $Q = 6, M = 10$ and $\zeta = 1/4$.

Figure 8 shows how the probability of degree distribution $P(k)$ for different values of M. They are very similar to those in Figure 2 obtained for actual texts. $P(k)$ goes through a maximum and then decreases according to a power law. We note that M is directly related to position of the maximum value, otherwise it has little influence on the value of the exponent γ. On the other hand, the model indicates that this maximum is due to the fact that networks grow by inclusion of cliques, with a characteristic size. This evolution dynamics favors the presence of nodes with a number of neighbors roughly given by maximal size of the cliques. In Figure 9 we indicate how the exponents γ and ζ are related. When $\zeta \to 0$, $\gamma \to 2.7$, being thus quite close to Barabasi's model. On the other hand, γ decreases when ζ increases and, for values ~ 0.25, it agrees quantitatively with the values obtained in the text analyses. For still larger values $\zeta \geq 0.5$ the scale-free behaviour is lost.

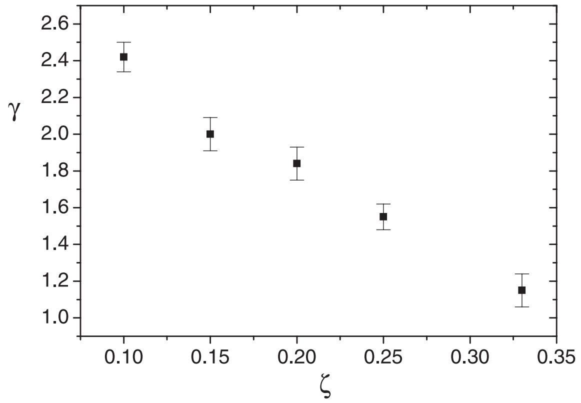


Fig. 9. Dependence of γ for different input values of ζ.

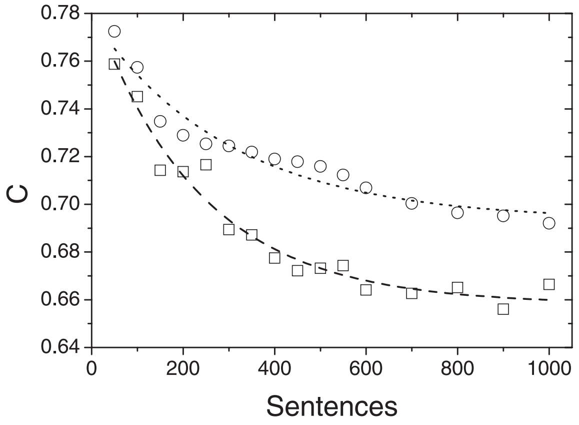


Fig. 10. Dependence of C as function of the number of sentences for two distinct values of M. The curves are exponential decay fits to the scattered points.

In Table 2 we indicate how the different indices used for network characterisation depend on M. The large values for the clustering coefficient (although actual texts have still larger values) indicates that this evolution model gives rise to networks with both small-world and scale-free features. In Figure 10 we show how C evolves with the number of cliques for different values $M = 10$ and 20, for networks with the same number of cliques $Q = 1000$. As in Figure 3, the points can be reasonably fitted by an exponential decay to an asymptotic value $C_{0,i}$.

It is clear that the limiting value of C depends on both Q and M, and we must take care in extrapolating results to the infinite limit. We call the attention that, with an average 60 000 words vocabulary, a much larger number of edges $\sim 10^{10}$ would be necessary to be in the situation of a complete graph. The average number of sentences in the text and model we analysed is much smaller than this limiting value, so that we are safely working in the sparse graph limit.

Table 2. Values of some networks indices for three simulated texts with 1000 sentences, $\zeta = 0.25$ and different maximum sentence sizes.

M	ℓ	C	$\langle k \rangle$	γ	k at $p(k)_{max}$
10	3.58 1	0.65	2.30	1.61	9
20	3.08 1	0.67	2.30	1.54	17
40	2.83 1	0.68	2.22	1.65	38

5 Conclusions

In this work we pursued the construction and analysis of networks the nodes of which are words with an intrinsic meaning in written texts. Edges were drawn among words that are present in complete sentences. The results obtained indicate that the networks have very robust properties, and the values for the network indices point to the presence of both small-world and scale-free features. This analysis differs from other previous works, in the sense that both words and sentences play an important role in the architecture of the network. We also characterized important aspects observed as the text grows, and studied the influence of distinct shuffling procedures of words and sentence size distribution.

The text investigation motivated us to investigate a model for network growth that was able to reproduce many of the features in the actual text networks. This model takes into account preferential attachment, addition of whole cliques (and not individual vertices) to the network, and a decrease of the probability of new words being in sentences that are added in latter stages of the network construction. With this ingredients, we did find results that are characteristic to scale-free networks, but with rather small values for the decaying exponent γ, as well as large values of the clustering coefficient C.

To conclude, we have briefly mentioned some efforts concerning the construction and analyses of networks that represent the relations between neurons in brain cortex [18]. On the other hand, analysed networks result from intense intellectual activity required to produce texts. The fact that networks present similar properties, e.g., both satisfy small-world and scale-free scenarios, may be of significance to bridge brain patterns to mind products, shedding light into the deeper problem of how the human mind works.

The authors would like to thank Fernanda Regebe, Gesiane Teixeira and Charles Santana for helpful discussions and assistance by code programming.

References

1. D.J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness* (Princeton University Press, 1999)
2. M. Buchanan, *Nexus: small world and the groundbreaking science of networks* (W.W. Norton & Company, Inc., New York, 2002)
3. A.L. Barabasi, *Linked: The New Science of Networks* (Perseus Books Group, Cambridge MA, 2002)
4. S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ. Press, 2003)
5. R. Albert, A.L. Barabasi, Rev. Mod. Phys. **74**, 47 (2002)
6. J. Camacho, R. Guimerà, L.A.N. Amaral, Phys. Rev. Lett. **88**, 228102 (2002)
7. A.-L. Barabási, R. Albert, H. Jeong, Nature **401**, 130 (1999)
8. J. Guare, *Six Degrees of Separations: A Play* (Vintage, New York, 1990)
9. D.J. Watts, S.H. Strogatz, Nature **393**, 440 (1998)
10. A.L. Barabasi, R. Albert, Science **286**, 509 (1999)
11. A.E. Motter, A.P.S. de Moura, Y.C. Lai, P. Dasgupta, Phys. Rev. E **65**, 065102 (2002)
12. L.F. Costa, Intl. J. Mod. Phys. C **15**, 371 (2004)
13. R.V. Solé, Nature **434**, 289 (2005)
14. R.F.I. Cancho, R.V. Solé, Proc. R. Soc. London, Ser. B **268**, 2261 (2001)
15. R.F.I. Cancho, R.V. Solé, R. Köhler, Phys. Rev. E **69**, 051915 (2004)
16. R.F.I. Cancho, R.V. Solé, R. Köhler, e-print `arXiv:cond-mat/0504154` (2005)
17. S.N. Dorogovtsev, J.F.F. Mendes, Proc. R. Soc. London, Ser. B **268**, 2603 (2001)
18. V. Eguiluz, G. Cecchi, D.R. Chialvo, M. Baliki, A.V. Apkarian, Phys. Rev. Lett. **92**, 018102 (2005)
19. G.K. Zipf, *Human behaviour and the principle of least effort. An introduction to human ecology* (Hafner, New York, 1972)
20. Unitex is a multilingual corpus processing system, based on automata-oriented technology developed by the Computational Linguistics Group of the Institut d'élétronique et d'informatique Gaspard-Monge (IGM-France) - `http://infolingu.univ-mlv.fr/english/`
21. Literary texts have been mostly obtained from Gutenberg Project website at `www.gutenberg.org`