International Journal of Modern Physics C
Vol. 21, No. 3 (2010) 333–347

DOI: 10.1142/S0129183110015142

COMPLEX SEMANTIC NETWORKS

G. M. TEIXEIRA
Centro Federal de Educação Tecnológica da Bahia
Salvador, Bahia, Brazil
gesianet@gmail.com

M. S. F. AGUIAR
Instituto de Física, Universidade Federal da Bahia
40210-340 Salvador, Bahia, Brazil

C. F. CARVALHO and D. R. DANTAS
Departamento de Psicologia, Universidade Federal da Bahia
40210-340 Salvador, Bahia, Brazil

M. V. CUNHA
Instituto de Física, Universidade Federal da Bahia
40210-340 Salvador, Bahia, Brazil

J. H. M. MORAIS
Departamento de Psicologia, Universidade Federal da Bahia
40210-340 Salvador, Bahia, Brazil

H. B. B. PEREIRA*
Programa de Modelagem Computational — SENAI CIMATEC
41.650-010 Salvador, Bahia, Brazil
Departamento de Ciências Exatas — UEFS
44031-460 Feira de Santana, Bahia, Brazil
hbbpereira@gmail.com

J. G. V. MIRANDA
Instituto de Física, Universidade Federal da Bahia
40210-340 Salvador, Bahia, Brazil
vivas@ufba.br

Received 4 January 2010
Accepted 26 January 2010

Verbal language is a dynamic mental process. Ideas emerge by means of the selection of words from subjective and individual characteristics throughout the oral discourse. The

*Corresponding author.

goal of this work is to characterize the complex network of word associations that emerge from an oral discourse from a discourse topic. Because of that, concepts of associative incidence and fidelity have been elaborated and represented the probability of occurrence of pairs of words in the same sentence in the whole oral discourse. Semantic network of words associations were constructed, where the words are represented as nodes and the edges are created when the incidence-fidelity index between pairs of words exceeds a numerical limit (0.001). Twelve oral discourses were studied. The networks generated from these oral discourses present a typical behavior of complex networks and their indices were calculated and their topologies characterized. The indices of these networks obtained from each incidence-fidelity limit exhibit a critical value in which the semantic network has maximum conceptual information and minimum residual associations. Semantic networks generated by this incidence-fidelity limit depict a pattern of hierarchical classes that represent the different contexts used in the oral discourse.

Keywords: Complex networks; oral discourse; incidence-fidelity index.

1. Introduction

In the last two decades, a broad set of transformations is highlighting the research and the understanding of language organization and language processes, particularly the interface among different language fields.[1] Currently, language studies are restricted by a wide range of combinations and paradigmatic breaks that emphasize the need of changing old disciplinary areas in order to gain a more comprehensive understanding of the mental processes and their relationship with the brain. Within this perspective, Martin *et al.*[2] used positron emission tomography (PET) and functional magnetic resonance imaging (MRI) to study the organization of semantic representations in patients with brain damage. One of the ways of expressing mental processes occurs through the acquisition and use of language.

Language represents and helps to build the self and subjectivity. Language is a very useful tool and can be used for the individual to understand, since it is able to retrieve information stored in memory. The individual speaks her/his thoughts through the use of words and symbols, thus expressing her/his knowledge, values and personal beliefs. The human linguistic system is characterized by complexity and articulation in a network of several neurological and cognitive processes.[3]

According to Pinker,[4] human language is structured from different formats of mental representations such as: images, phonologic links, hierarchical and mental trees. For the author, the structure of these components is organized according to the logic of complex systems and inherent biological artifacts.

This engagement in the study of complex properties of the language and of its configuration as an important element of cognition, focuses on fields devoted to the understanding of the language. The main purpose for that is to develop models capable of representing and explaining a set of processes previously avoided by researchers from the field (e.g. intentionality, metaphorical capacity, coordination with the emotions, formations of concepts and subjectivity). As a result, interdisciplinary scientific fields appear in order to describe and analyze languages. Within this context, language is understood as a set of highly complex systems which

works with multiple levels of organization that include integrated processes related to emotions, reason and biological regulation.

Cognitive Sciences represent one of these interdisciplinary scientific fields and insert semantic networks as theoretical foundations of the study of cognitive processes. Semantic networks are used as (1) graphic representations for simulating knowledge and/or (2) support tool for automated systems of inferences about knowledge from a network structure.[5] Thus, the network simulates the relationship between objects and their codes, offering a survey through graphic mapping.

From the beginning, traditional models of semantic networks applications have emerged from the areas dedicated to the study of language and its processes: for instance, Linguistics[6]; Cognitive Sciences[7]; Neurosciences[8] and Artificial Intelligence.[9] Semantic memory is responsible for the organization of knowledge of words and other symbols. Thus, semantic memory works with concepts that enable other ideas associated to them.[10] Symbols, concepts and relations of the semantic memory mainly represent the networks of association within semantic networks. The goal of this paper is to present a model for building a linguistic network based on the semantic network of oral discourse. Complex network models are used to extract semantic networks from individual oral discourses.

2. Building a Semantic Network

A non-directed network is defined as a set of nodes, called vertices, some of which are connected by lines called edges. In a formal way, a graph $G = (V, E)$ is a mathematical structure that consists of two sets: V (finite and not empty) and E (binary relation on V). The elements of V are called vertices and the elements of E are called edges. Each edge has a set of one or two vertices associated to it.[11] However, in our semantic networks, we take into account only that each edge has two vertices associated to it. This definition can be used to characterize a large number of systems such as social networks,[12] scientific collaboration and biological systems,[13] co-authorship,[14,15] webs of words,[16] quotations,[17] information networks within organizations,[18] World Wide Web[19] among others. The topological structures of networks of natural systems have, in most cases, patterns of organization with characteristics of complex systems.

To build a semantic network based on oral discourse, we have preserved only the words with an intrinsic meaning, called content words. Words that merely have grammatical functions related to the arrangement of syntactic structures of sentences in the text (articles, pronouns, prepositions, connectors, abbreviations, and interjections), also called function words, have been eliminated. In order to perform a computer implementation, we have used some routines, dictionaries, and grammatical rules from the UNITEX package.[20] A detailed description of the text treatment has been presented in Ref. 21.

We hypothesize that words that occur together in the same sentence would have been evocated in an associative way for the building up of the idea to be presented.

This is the central idea in the network construction. So we can build a network where the words are represented as the vertices, and an edge is created to connect pairs of words that occur in the same sentence.[21] However, using this criterion, word pairs whose association is not very significant, are included in the network and mask the structure formed by the stronger associations. It is necessary to provide a filter so that only the most relevant associations for the discourse are considered in the network construction. Previous works on semantic network construction from surveys[22] considered the frequency of a pair of words, called force of association, an important criterion of word associations.

Nevertheless, in the case of the oral discourse, we believe that this criterion is not enough. In the oral discourse context there is a tendency for the frequency of a pair of words to be more related to the content of the discourse than the intrinsic semantic association between the concepts represented by the words. Therefore, we propose an index that we call "incidence-fidelity index". This new index takes into account not only the frequency of occurrence of a pair of words (i.e. concept of force as defined by Nelson *et al.*[22]) in the oral discourse, but also their probability of occurring together (i.e. concept of fidelity).

2.1. *Incidence-fidelity concept*

Nelson *et al.*[22] developed the concept of force between pairs of words, measuring the frequency of association or probability of a word be linked to another, using the technique of discrete free association. In Nelson *et al.*'s paper[22] the concept of force is defined as the ratio between the frequency in which a pair of words occurs and the total number of times the given word is found. A large data base produced by Nelson and colleagues was used by Ref. 23 to evaluate the topological properties of semantic networks built from the force of pairs of words. The network represented the complex mechanism of association between concepts that emerged from individuals. However, how representative is this network? As mentioned earlier, within the discourse of an individual, the context of the discourse plays a role. Consequently, the frequencies of words depend on the discourse topic. To minimize this effect, we propose the concept of incidence-fidelity index.

The index formed by the concepts of incidence and fidelity must take into account not only the frequency of occurrence of a pair of words (i.e. definition of force by Nelson *et al.*[22]), but also their probability of co-occurrence. This is what we call the "fidelity" of the pair of words. The "incidence concept" is exactly the same as "force" defined by Ref. 22. Here we use the name incidence in order not to cause misleading with the physical meaning of force in the Newtonian concept. The goal of the Incidence-Fidelity index (IF) is to merge both concepts of incidence and fidelity. That is, the IF index represents the probability of occurrence of the pair of words in the oral discourse context as a whole and the probability of the words of the pair always occurring together.

The concept of incidence in this work is defined as the normalized frequency in which a pair of words occurs in an oral discourse. And the concept of fidelity is defined as the probability of a pair of words always occurring together. The concept of incidence-fidelity (i.e. *IF* index) is the union of these two definitions.

This concept is best explained through by the theory of sets. The set of sentences of oral discourse in which a certain word w appears is defined as C_w. In our case, given a pair of words defined by variables ϕ and ψ, we can define the sets C_ϕ and C_ψ, where C_ϕ is the set formed by sentences in which the word ϕ occurs and C_ψ is the set formed by sentences in which the word ψ occurs. Thus we can define the subset of sentences in which the words co-occur in a sentence by $C_p \equiv C_\phi \cap C_\psi$ and its cardinality correspondent as $S_p \equiv |C_\phi \cap C_\psi|$.

From these definitions and using the theory of sets as a starting point, we can define the concept of incidence, namely:

$$I \equiv \frac{|C_\phi \cap C_\psi|}{|\bigcup_{i=1}^{N_p} C_i|}\,, \tag{1}$$

where N_p refers to the total number of words of the oral discourse, so that $\bigcup_{i=1}^{N_p} C_i$ calculates the total number of sentences of oral discourse. The incidence I represents the probability of the subset C_p to occur within the universe of possibilities of the whole oral discourse.

The concept of fidelity is defined as:

$$F_i \equiv \frac{|C_\phi \cap C_\psi|}{|C_\phi \cup C_\psi|}\,. \tag{2}$$

This represents the probability of occurrence of the pair of words within the universe of possibilities of words of the pair.

Using the concepts of fidelity and incidence, F_i and I respectively, as a starting point, we present the concept of incidence-fidelity defined as the product between them:

$$IF = \frac{S_p^2}{N_S(S_\phi + S_\psi - S_p)}\,, \tag{3}$$

where N_S is the total number of sentences in the text.

In Eq. (3), we note that index *IF* can assume values between 0 and 1. Zero means that the pair of words never occurred together and one means that all sentences of text contain the pair of words.

2.2. *Origin of data*

The transcriptions of oral discourses of 12 individuals (i.e. undergraduate students of Physics and Psychology) were used as main data to carry out experiments and to verify and validate the methodology proposed. These students were interviewed in order to produce approximately one hour long free oral discourses, where the main subject was "*I*". Psychologists who carried out the interviews interfered as

minimally as possible, always using words that were within the oral discourse of the interviewee. This procedure is used to minimize possible suggestions of new association routes. It is important to comment that all oral discourses were transcribed by the same researcher.

2.3. *Resulting semantic networks*

The values of the fidelity, incidence and incidence-fidelity indices (F_i, I, and IF, respectively) were calculated taking into account all pairs of words of each oral discourse transcription. Typical distributions of F_i, I, and IF values for an oral discourse are depicted in Fig. 1. The occurrence frequency of I and IF values in all oral discourses has a power law probability distribution. This behavior shows that the oral discourses are mostly structured by not very meaningful word pairs and that there is a small core of associations containing pairs of words with high I and IF values. On the other hand, F_i behaves differently. The high occurrence of high F_i values indicates that there are pairs of words that appear only once in the discourse. However, these pairs of words present in almost all cases low I values (this effect can be observed in the inset of Fig. 1). This compensation between F_i and I indices are incorporated in IF index [Eq. (3)].

The method of constructing the semantic network proposed in this paper is based on the construction of an association network for pairs of words using IF values as a criterion in order to filter the most significant associations of the text; that is, an edge is removed between the pairs of words having an IF value lower than a given limit IF value (IF_L), and when the vertex has no edges it is also removed. Taking into account occurrence frequency and co-occurrence probability

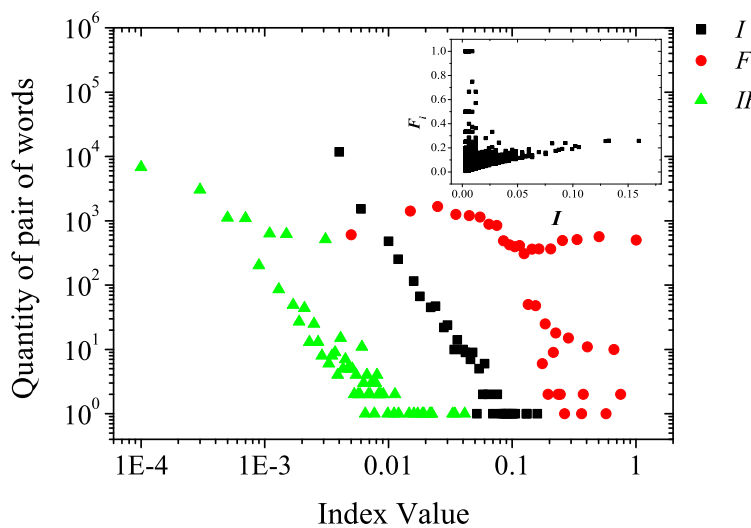
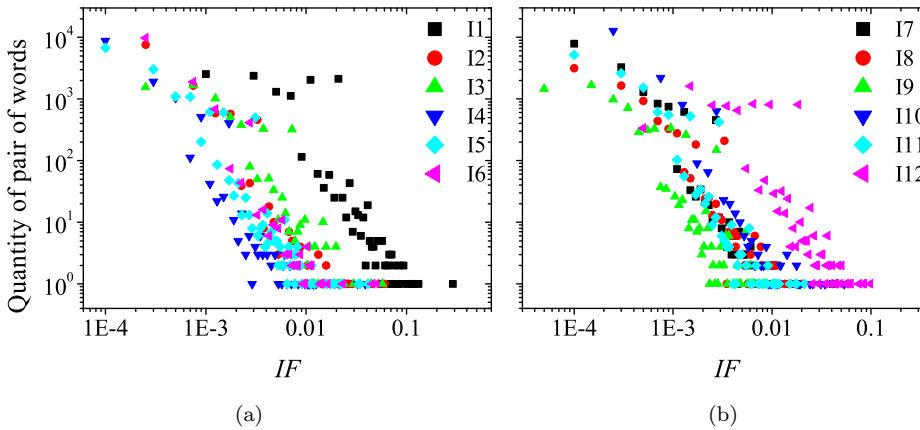


Fig. 1. Distribution of IF values of the oral discourse $I5$. The fitted line has slope $\gamma = -1.8$, $\sigma = \pm 0.1$.



(a) (b)

Fig. 2. Distributions of *IF* values of 12 oral discourses. Distributions have an average slope $\langle\gamma\rangle = -1.9$, $\sigma = \pm 0.09$.

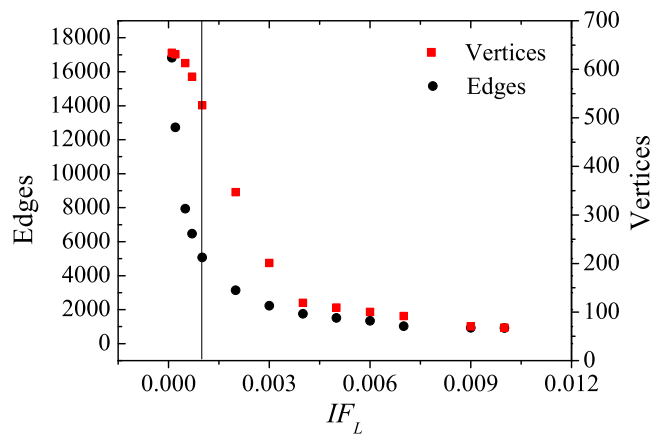
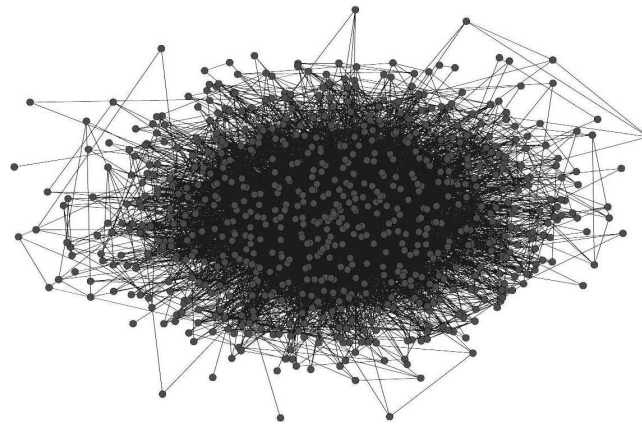


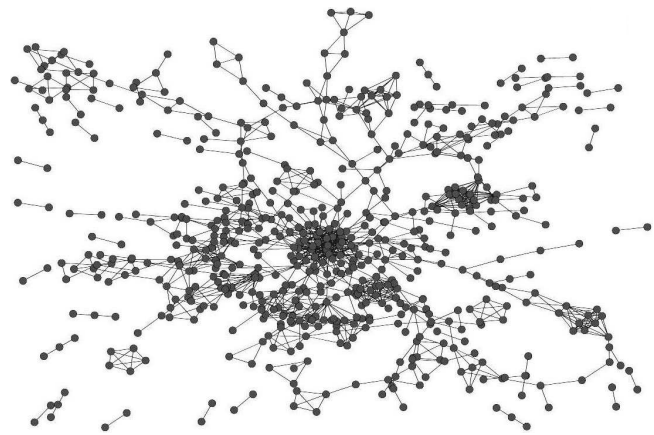
Fig. 3. The average number of vertices and edges of 12 oral discourse networks as a function of IF_L value.

we hope to select only the most significant associations of the text. According to the 12 distributions shown in Fig. 2, the quantity of associations is quite sensitive to the IF_L value. That is how we calculate the quantity of associations (edges) and the number of words (vertices) for different IF_L values (Fig. 3).

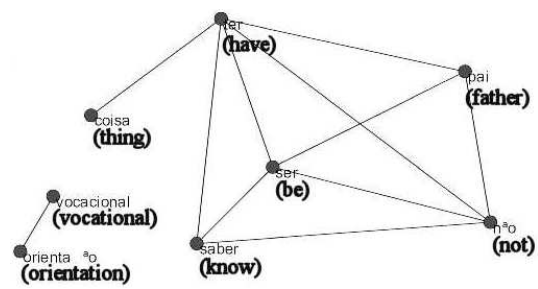
The curves shown in Fig. 3 can be interpreted as the quantity of information that is eliminated when the IF_L value is increased. We observed that the number of edges of the oral discourse network decreases faster than the number of vertices. In addition, for $IF_L \sim 10^{-3}$ the derivate (i.e. fall velocity) is maximum for the number of words (vertices). This value ($IF_L \sim 10^{-3}$) represents the minimum necessary quantity of edges of the network while maintaining a maximum number of vertices; we call this the Critical Incidence-Fidelity (IF_C) value.



(a)



(b)



(c)

Fig. 4. Three semantic networks generated from the oral discourse of the individual $I11$ taking into account three possible IF_L values, i.e. (a) $IF_L = 10^{-4} < IF_C$; (b) $IF_L = IF_C$ and (c) $IF_L = 10^{-2} > IF_C$, for $IF_C = 10^{-3}$.

Figure 4 depicts an oral discourse network of one of the individuals. As observed, the three networks generated depend on the different IF_L values ($IF_L < IF_C$, $IF_L = IF_C$ and $IF_L > IF_C$). The premise of this method is that the associations with IF_L value smaller than IF_C value do not belong to the core of more significant associations of the oral discourse [Fig. 4(a)]. On the other hand, an oral discourse network with IF_L value bigger than IF_C value loses much information (i.e. words) and this represents a limitation in the use of oral discourse network to establish a semantic relation [Fig. 4(c)].

In Fig. 4, we observe a typical behavior of change of network topology in function of the IF_L value. In a general way, we observe that there is a significant change of the structure of connection of the words for the different IF_L values. We use indices from the graph theory and complex networks to better characterize network topology.

3. Characterization of the Semantic Networks

Networks of word associations have a typical complex networks behavior.[16,23] Because of that we use the basic statistical indices from complex network theory to characterize this kind of networks. The indices used are diameter (D), mean distance (l), clustering coefficient (C) and degree distribution ($P(k)$) of the network. All indices were calculated for different IF_L with the exception of degree distribution because it is for the network to have a sufficient quantity of vertices in order to observe a pattern in the degree distribution (this does not occur for high values of IF_L, see Fig. 3).

The indices of discourse $I5$ semantic networks are presented in Figs. 5, 6 and 9, generated for increasing values of IF_L. D and l indices are associated to the

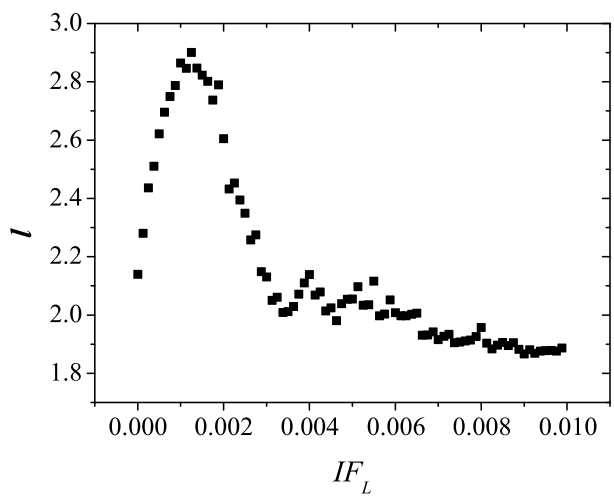


Fig. 5. Values of the mean distance (l) for increasing values of IF_L of the oral discourse $I5$.

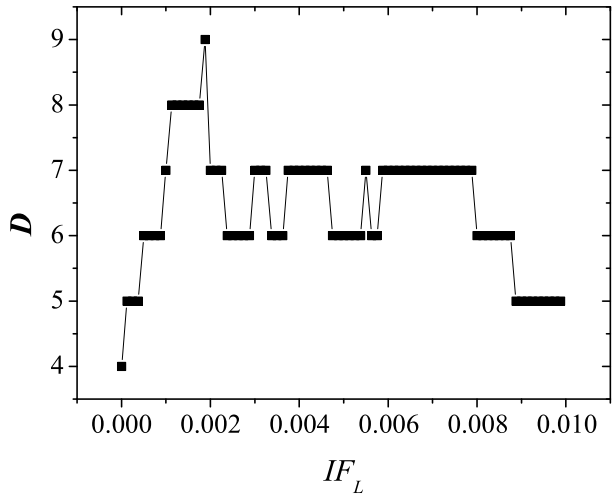


Fig. 6. Diameter values calculated from the semantic network of the oral discourse $I5$ for increasing values of IF_L.

distances between the words in the semantic network, where D is the maximum distance and l is the mean distance. l is obtained from the calculation of the average value of the minimal path between vertices in the network and the diameter D is the supremum over the set of minimal paths between all the vertices.[13] For increasing values of IF_L, initially we observe a growth of l values, reaching a maximum value for IF_L values close to IF_C. After that, we observe a fast decrease of l values and a subsequent stabilization of l values. The existence of a maximum value of l, when $IF_L = IF_C$, suggests that the semantic network for this value has the widest range associations of the oral discourse [Fig. 4(b)].

It is important to notice that if we use only the incidence index,[22] the network indices decrease monotonically without any critical value. This is expected due to the fact that the incidence index represents only the probability of finding a pair of words in the text and as already indicated by Zipf[24] the distribution of this type of index decreases monotonically. In Fig. 7, we present an example of this behavior of one of the oral discourses.

There are several definitions of clustering coefficient. For Watts and Strogatz,[25] clustering coefficient "measures the cliquishness of a typical neighborhood (a local property)" and for Newman[26] clustering or transitivity, from a network topology perspective, "means the presence of a heightened number of triangles in the network — sets of three vertices each of which is connected to each of the others." We used the definition presented by Albert and Barabási,[27] "the ratio between the number E_i of edges that actually exist between these k_i nodes and the total number $k_i(k_i - 1)/2$ gives the value of the clustering coefficient of node i" [Eq. (4)]:

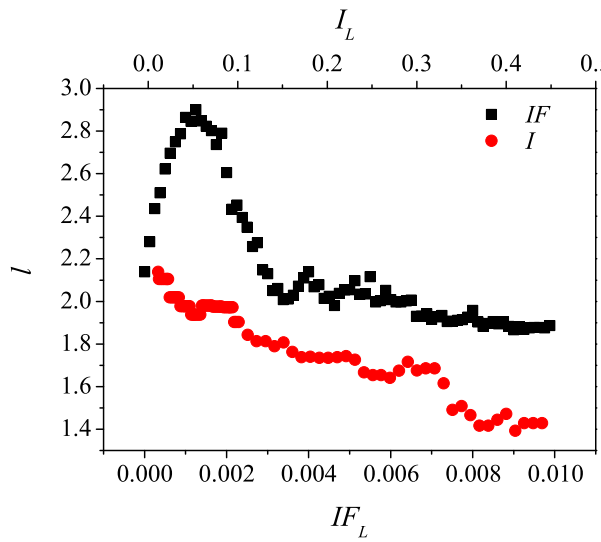


Fig. 7. Comparison between utilizing the *IF* index and only the incidence index for the oral discourse *I*5.

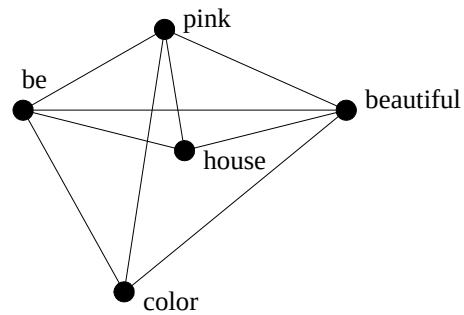


Fig. 8. A network of word associations from the following sentence: "The house is beautiful. The house is pink. The pink color is beautiful!"

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \tag{4}$$

where C_i is the clustering coefficient of vertex i, E_i is the number of edges between the neighbors of the vertex i and k_i is the number of connections or degree of vertex i (i.e. number of incident edges taking into account an undirected network).

The clustering coefficient of a vertex measures the fraction of connections between neighbors of a vertex that are connected to each other. From the network of word associations, this index characterizes the degree of mutual associations between words selected during the oral discourse (e.g. Fig. 8).

In the example of Fig. 8, the clustering coefficient of the word "house" is 1 (one), because there are connections between all words linked to it.

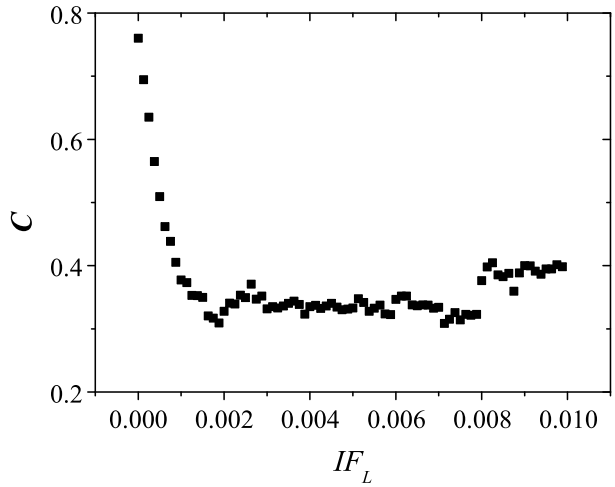


Fig. 9. Values of clustering coefficient C for different values of IF_L of the oral discourse $I5$.

The clustering coefficient of the network (C) is calculated from the average of all individual C_i. Figure 9 shows the values of C in function of IF_L values. In Fig. 9, we observe that the value of clustering coefficient for $IF_L = IF_C$ is equal to 0.58. This high value of clustering coefficient, compared with a similar random network (i.e. the same number of vertices and average degree $\langle k \rangle$), and the diameter value (D) of 16 (four times larger than its mean distance l) suggest to us that the topology of the critical network presents modular characteristics, where the groups of words are more strongly linked forming modules connected by weak ties. A stronger proof of the modularity is by using $C(k)$ and assortative index, however the number of vertices of the critical networks (Fig. 3) are not sufficient to estimate such indices. Due to this modular characteristic, the averages of clustering coefficient and mean distance of the network decrease. Furthermore, the connections between the modules by weak ties increase the network diameter.

The topology obtained from these results suggests a categorized structure of oral discourses. When we zoom in the modules, we observe that each module can be seen as different contextualized instances of the oral discourse (Fig. 10).

4. Concluding Remarks

As an approach to the complex behavior of language, we have elaborated a method able to generate a contextualized semantic network from the oral discourse of an individual. This method uses concepts and properties of Complex Networks and Set Theory to identify the network that best represents the structure of word associations of a discourse.

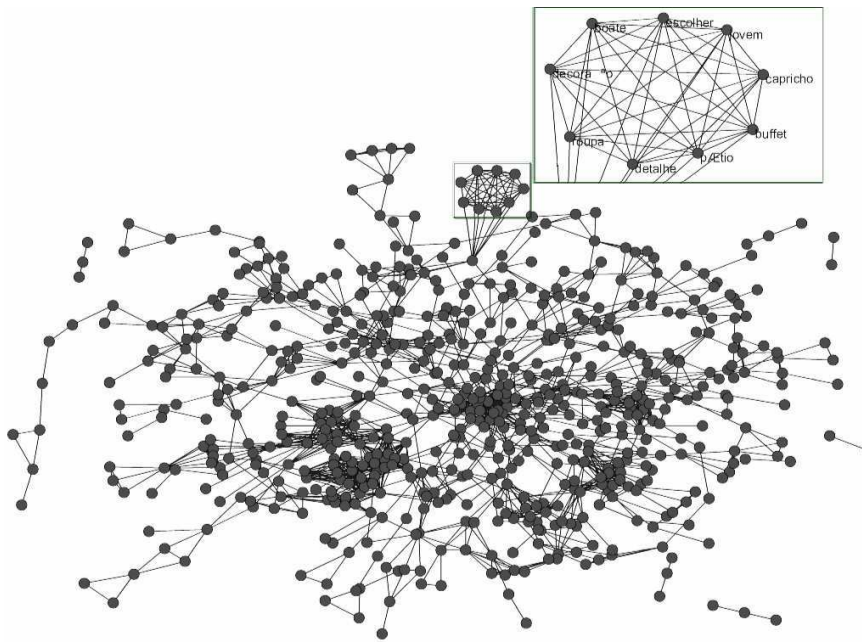


Fig. 10. A critical network. The detail in upper right-hand corner represents the zoom in on a subnetwork from the oral discourse of Individual 2.

The frequency distribution of pairs of words in the semantic networks of the oral discourses studied follows a power law. This behavior has already been observed by Zipf.[24]

The indices of networks obtained from different values of IF_L show a typical behavior of phase transition with a very well defined critical point. The network generated at this critical point is called a critical network and is used to represent the characteristic network of the oral discourse. The average of IF_C value for all oral discourses was 10^{-3}. The critical networks generated with $IF_C = 10^{-3}$ show values of diameters (D) and clustering coefficients C much larger than random networks with the same size, and this is related to a modular topology with clusters linked by weak ties. For $IF_L < IF_C$, the small D and high C show a very dense network. On the other hand, for $IF_L > IF_C$, we get a sparse network since both indices D and C are small. In both cases the networks are not modular.

From the complex network perspective, all the networks studied showed topologies with characteristics of small-world networks. In addition, for some networks with greater number of vertices, the degree distributions present characteristics of a power law, suggesting a topology of scale-free networks. In order to verify this, it would be necessary to examine other networks with a high number of vertices, but this would require studying longer oral discourses.

The property of complex networks used in the semantic networks proposed in this work shows results similar to Steyvers and Tenenbaum's work[23] on complex

networks and semantic networks. Both works also have similar topologies for the semantic networks. Despite this similarity, the structure of our semantic networks is modular and has high values of diameters (D) and clustering coefficients C. This does not happen in the semantic network of word association found in Ref. 23.

Although there is some variability in the value of IF_C for the different oral discourses analyzed, all discourses presented the same critical behavior with a well defined critical network and with similar topologies. This indicates the possibility that such behavior and critical topology are intrinsic characteristics of the mechanism of the human language. Finally, semantic networks from oral discourses could contribute to the development of new methods for psychometric research (e.g. identification of schizophrenic people since their oral discourses suggest semantic networks for $IF_C = 10^{-3}$ without a semantically logical backbone, but characterized by some unconnected islands).

Acknowledgments

We would like to thank the undergraduate students of Physics and Psychology of the Federal University of Bahia, the undergraduate students of Psychology of the Bahia School of Medicine, the FESC Group (Computational Statistics Physics) of the UFBA, the CONES Group (Modeling on complexity Neuroscience Art and Health) of the UFBA, Claudio Silva, Denise Coutinho and Silvia Caldeira.

This research has been partially supported by the FAPESB (Bahia State Grant Agency) under the project number CO-112/2005.

References

1. M. S. Gazzaniga, R. B. Ivry and G. R. Mangun, *Cognitive Neuroscience: The Biology of the Mind* (W. W. Norton and Company, 2008).
2. A. Martin, C. L. Wiggs, L. G. Ungerleider and J. V. Haxby, *Nature* **379**, 649 (1996).
3. A. Damásio, *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Harcourt, New York, 2000).
4. S. Pinker, *The Language Instinct* (William Morrow, New York, 1994).
5. J. F. Sowa, *Semantic Networks* (2002).
6. A. Caramazza, *Nature* **380**, 485 (1996).
7. A. M. Collins and E. F. Loftus, *Psychological Rev.* **82**, 407 (1975).
8. H. Damásio, T. J. Grabowski, D. Tranel, R. D. Hichwa and A. R. Damásio, *Nature* **380**, 499 (1996).
9. N. J. Farah and J. L. Mclelland, *J. Exp. Psychol. Gen.* **120**, 339 (1991).
10. R. J. Sternberg, *Cognitive Psychology* (Harcourt Brace College Publishers, London, 1999).
11. J. Gross and J. Yellen, *Graph Theory and its Applications* (CRC Press, Boca Raton, 1999).
12. D. J. Watts, *Small Worlds: The Dynamics of Networks between Order and Randomness* (Princeton University Press, Princeton, 1999).
13. A. L. Barabási and R. Albert, *Science* **286**, 509 (1999).
14. M. E. J. Newman, *Phys. Rev. E* **64**, 016131.1 (2001a).
15. M. E. J. Newman, *Phys. Rev. E* **64**, 016132.1 (2001b).

16. R. Ferrer and R. V. Solé, *Proc. R. Soc. London* (2001).
17. S. N. E. Dorogovtsev and J. F. F. Mendes, *Proc. R. Soc. London, Ser. B* **268**, 2603 (2001).
18. L. López, J. F. F. Mendes and M. A. F. Sanjuán, *Physica A* **316**, 695 (2002).
19. M. Thelwall and D. Wilkinson, *J. American Soc. Info. Science Technology* **54**, 706 (2003).
20. S. Paumier, UNITEX 2.0 User Manual, Electronic version, www-img.univ-mlv.fr/~unitex/UnitexManual2.0.pdf (2008).
21. S. M. G. Caldeira, T. C. Petit Lobão, R. F. S. Andrade, A. Neme and J. G. V. Miranda, *Eur. Phys. J. B* **49**, 523 (2006).
22. D. L. Nelson, C. L. Mcevoy and T. A. Schreiber, *The University of South Florida Word Association Norms* (1999).
23. M. Steyvers and J. B. Tenenbaum, *Cognitive Science* **29**, 41 (2005).
24. G. K. Zipf, *The Principle of Least Effort: An Introduction to Human Ecology* (Hafner Publishing Company, New York, 1972).
25. D. J. Watts and S. H. Strogatz, *Nature* **393**, 440 (1998).
26. M. E. J. Newman, *SIAM Review* **45**, 183 (2003).
27. R. Albert and A. L. Barabási, *Rev. Mod. Phys.* **74**, 49 (2002).