Detecting order-disorder transitions in discourse:

implications for schizophrenia

Álvaro Cabana[1], Juan C. Valle-Lisboa[1], Brita Elvevåg[2]*, Eduardo Mizraji[1]

1. Group of Cognitive Systems Modeling, Biophysical Section. Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay.

2. Psychiatry Research Group, Department of Clinical Medicine, University of Tromsø, & Norwegian Centre for Integrated Care and Telemedicine (NST), University Hospital of North Norway, Tromsø, Norway.

* Corresponding author.
Brita Elvevåg
UNN-Åsgård, 9291 Tromsø, Norway.
Email: brita@elvevaag.net

Abstract

Several psychiatric and neurological conditions affect the semantic organization and content of a patient's speech. Specifically, the discourse of patients with schizophrenia is frequently characterized as lacking coherence. The evaluation of disturbances in discourse is often used in diagnosis and in assessing treatment efficacy, and is an important factor in prognosis. Measuring these deviations, such as “loss of meaning” and incoherence, is difficult and requires substantial human effort. Computational procedures can be employed to characterize the nature of the anomalies in discourse. We present a set of new tools derived from network theory and information science that may assist in empirical and clinical studies of communication patterns in patients, and provide the foundation for future automatic procedures. First we review information science and complex network approaches to measuring semantic coherence, and then we introduce a representation of discourse that allows for the computation of measures of disorganization. Finally we apply these tools to speech transcriptions from patients and a healthy participant, illustrating the implications and potential of this novel framework.

Keywords: Discourse trajectory; incoherent speech; complex networks; topic entropy, schizophrenia.

1 Introduction

Language production and comprehension provide a window into the cognitive and neural architecture underlying complex information processing in the brain (Pinker, 2000). They are high-level cognitive functions that reflect the state of numerous cognitive processes. The pattern and content of the communication can be traced back to individuals' cognitive abilities, knowledge, affective state and consequently their overall mental state. Disturbances in the domain of language, especially in speech, occur in a variety of psychiatric and neurological conditions, and their neural substrates are likely to be related to the pathophysiology of the disorder (DeLisi, 2001), and hence are a fundamental aspect in diagnosis and assessing treatment responsiveness and prognosis (Andreasen and Grove, 1986; Andreasen and Black, 2005; McKenna and Oh, 2005).

Indexing language comprehension and production disturbances has been conducted using a variety of neuropsychological measures and tests (Hodges et al., 1992, McKenna et al., 1994 Tamlyn et al., 1992). We focus on speech, which traditionally has been quantified for predictability and variability using a variety of manual (and labor-intensive) techniques, such as cloze analysis, type:token ratios, analysis of lexical and syntactic structure, and also discourse structure using cohesion analysis (for a review, Kuperberg, 2010). There are a variety of fine-grained rating scales of the coherence of speech and communication, such as the Scale for the Assessment of Thought, Language and Communication (TLC; Andreasen, 1986), the Communication Disturbances Index (Docherty, 2005), and the Thought Disorder Index (TDI) (Solovay et al., 1987; Niznikiewicz et al., 2002), use of which requires extensive training but nonetheless remains open to variance across raters. In some sense these are probing "communication efficiency", which can be assessed by a range of computational linguistic techniques (Jurafsky and Martin, 2000). Indeed, such studies - using Latent Semantic Analysis (LSA) which models and matches discourse content (Landauer and Dumais, 1997; Landauer et al., 2007) - have demonstrated that it is possible to evaluate patients with schizophrenia based on open-ended verbalizations. These automatically derived language scores have distinguished patients from controls accurately (and patients from other patients, and also from their family members), using both large discourse samples as well as responses consisting of only a few words (Elvevåg et al., 2007, 2010).

Our goal here is to present some tools derived from recent developments in network theory and information sciences that enable the capture and indexing of "meaning" in a quantifiable and biologically relevant manner. This is because there are statistical properties in expressed language that provide a rich source of information regarding “meaningful communication”. Specifically, we present measurements of disorganization of discourse based on topic randomness and semantic graph measures. Thus, in the next section we describe methods based on information science and complex network approaches to language, and we introduce a particular representation of discourse, and present ways to measure its disorganization. Then we apply our framework to speech samples from patients with schizophrenia and a healthy participant to illustrate the potential of the method.

2 Methods

2.1 Semantic graphs and complex networks

Our method requires a simple but rich representation of meaning. One approach to achieve this is graph representation, with roots in semantic network theory (Collins and Loftus, 1975; Collins and Quillian, 1969; Steyvers and Tenenbaum, 2005). Graphs are mathematical objects consisting of sets of nodes and sets of edges connecting the nodes. Traditional semantic networks are “graphs” with labeled connections that instantiate different relationships between entities (e.g., “a robin is a bird” is represented by a particular type of link (the IS-A link) between the “robin” node and the “bird” node, or the HAS linking “a bird has feathers”, that together support the inference that “a robin has feathers” (Quillian, 1968)). Semantic graphs (i.e., “stripped down” versions of semantic networks) can be used to capture associative and conceptual relationships by automatically analyzing large portions of text, usually linking together nodes that represent words that co-occur within a small range in a large corpus (graphs built this way are referred to here as “lexical graphs”; Ferrer i Cancho and Solé, 2001; Dorogovtsev and Mendes, 2001; Steyvers & Tenembaum, 2001).

Recent developments in graph theory applied to the study of complex systems have shown that many natural and artificial complex networks show the small world and scale free properties (Albert and Barabási, 2002). The former means that networks tend to have high clustering coefficients [Footnote 1] while keeping low path lengths (Watts and Strogatz, 1998), and the latter implies that link distribution is frequently characterized by a power law, enabling highly connected nodes to appear relatively often. These two characteristics confer interesting properties to networks, like fast transmission and failure tolerance (Motter et al., 2002; Ferrer i Cancho et al., 2005; Steyvers and Tenenbaum, 2005). In order to represent meaning, we use graphs that turn out to satisfy some of these properties.
As our goal here is to capture the thematic structure of a single instance of a linguistic

expression (a relatively small sample of text, discourse or dialog) we will represent words as nodes in a semantic graph, and consider discourse as a trajectory in such a graph. “Goal directed” discourse would show an ordered and organized trajectory, whereas thought disordered discourse would appear as a disordered trajectory due to the disorganization of the semantic structure, or of the mechanism that searches through it. We propose that to measure this disorganization, semantic structures be represented using networks and characterized using measures inspired by information theory. However, in order to derive useful tools two methodological challenges have to be addressed: First, find a suitable representation of discourse, including a topic graph and a trajectory. Second, devise measures of disorganization, sensitive enough to detect subtle deviations.

2.2 The representation of discourse

To represent discourse we analyzed textual transcriptions of speech samples. The texts were subjectively evaluated by delimiting small blocks of text of just one theme or idea, and labeling each block in the text with a set of words representing that theme (Cabana, 2009). As a calibration procedure we analyzed the first two chapters of “A study in scarlet” by Arthur Conan Doyle, as the descriptive nature of the text is devoid of complex metaphors or other literary devices that could complicate analysis. For this large text sample, we added the additional criteria that block size should be between two sentences and four paragraphs and that the selected theme be distinguishable from the previous and next blocks. We show this calibration example in figure 1. As the discourse advances, consecutive paragraphs share some labels, allowing the construction of a graph, whereby each label is a node and each pair of labels that co-occur in one block of text is linked by an edge (figure 1 B). The resulting graph is the topic graph (figure 1 C).

[PLACE FIGURE 1 AROUND HERE]

When the topic graph is displayed bi-dimensionally, the discourse trajectory can be represented as a line drawn over each block in the order of their appearance in the discourse (figure 1 B, D). If this trajectory is drawn over the whole topic graph, the line appears convoluted and folded, as a result of the text “re-visiting” central topics of the story (figure 1D). These “long range interactions” are what we seek to capture by measuring the entropies (see Section 2.3). The rationale for this *a priori* expectation is that sentence sequence in discourse is not random when language is organized (see Kintsch, 1988; Foltz, 2007), and loss of this higher-level order would result in disorganization. From visual inspection of the topic graph and the trajectory line, at least five major sets of labels can be identified that delimit five major topics in the graph (table 1). In a larger graph, this delimitation could be automatically performed by identifying connected components (sets of connected nodes disconnected from others) or communities (sets of nodes statistically more connected to each other) as representing major topics (Palla et al., 2005; see Section 4).

[PLACE TABLE 1 AROUND HERE]

2.3 Measures of disorganization

The central hypothesis of this work is that loss of goal, tangentiality and incoherence frequently observed in schizophrenia are based in part on problems "following" an ordered trajectory among different topics. In coherent discourse adjacent words refer to connected topics. In contrast, in incoherent discourse a certain degree of "shuffling" of the topics occurs such that adjacent words may belong to different topics. This does not imply a "word salad" as the discourse may respect syntax, word order and even word similarity, but nevertheless reveal a high degree of disorder - semantic shuffling - in terms of meaning. To detect and measure this disorganization, we developed topic and transition entropy measures (closely related to that used in statistical mechanics).

2.3.a Topic entropy

T*opic entropy* $S(\alpha)$, is defined as

$$S(\alpha) = -\sum_{i=1}^{n(\alpha)} p(\alpha_i) \log p(\alpha_i) \qquad (1)$$

α being a particular topic in the text, $n(\alpha)$ the number of continuous stretches of text attributable to this topic and $p(\alpha_i)$ the ratio of length of the stretch i to total number of words attributable to this topic. This equation (1) measures the level of discontinuity of words belonging to the same topic. When discourse is organized in a perfect sequence of topics each consisting of an uninterrupted stretch of text, all topic entropies will be zero, and will grow in those cases where topics are more interspersed. To illustrate, we shuffled the text of our example, "A study in scarlet", and present a visual representation of this shuffling (figure 2).

[PLACE FIGURE 2 AROUND HERE]

Figure 3 A shows how the entropy of each of the topics increases as a result of the shuffling. The first instances of shuffling disrupt the original order only moderately, and a great deal of

shuffling has to be imposed to result in some “randomness”, but the greatest increases in entropy occur in the first shuffling instance. Given that every topic’s entropy increases, but to different degrees, a good measure of disorganization is mean entropy across topics, estimated as

$$\overline{S(D)} = \frac{1}{N_T}\sum_{\kappa}^{N_T} S(\kappa) \qquad (2)$$

where $\overline{\mathrm{S(D)}}$ refers to the entropy of the discourse and $\mathrm{N_T}$ is the number of topics that are expressed.

2.3.b Transition entropy

Transition entropy, is defined as

$$T(\alpha) = -\sum_{\tau \neq \alpha} p_\alpha(\tau)\log p_\alpha(\tau) \qquad (3)$$

$\mathrm{p}_\alpha(\tau)$ is the fraction of transitions from topic α to topic τ. Discourse can have a large topic entropy (calculated with equation (1)) but zero transition entropy (calculated with equation (3)). If the discourse were perfectly periodic (e.g., a repetition of sequence $\alpha\beta\gamma\alpha\beta\gamma\cdots$), then transition frequencies would be $\mathrm{p}_\alpha(\beta)=1$ and $\mathrm{p}_\alpha(\gamma)=0$, and the entropy defined by equation (3) is equal to 0. With reference to figure 3 B illustrating transition entropy for the original and shuffled versions of “A study in scarlet”, the increase in transition entropy is apparent, and the increase is more subtle than with topic entropy. This is probably because in the original text the entropy is already high since topics are relatively independent, or because the small sample of transitions cannot be used to detect inter-topic structure. As with topic entropy, mean transition entropy can be defined as:

$$\overline{T(D)} = \frac{1}{N_T}\sum_{\kappa}^{N_T} T(\kappa) \qquad (4)$$

where the sum is performed over all topics.

[PLACE FIGURE 3 AROUND HERE]

It remains to be established whether subtle disorganization in semantic structure of discourse can be detected reliably and reproducibly using this approach. We present below examples of its potential usefulness.

3 The topology of speech in schizophrenia

Clearly our method would benefit from further refinements, but we nonetheless illustrate the potential usefulness of the whole approach and demonstrate its “proof of concept”.

One important difference between short speech transcriptions and the example we used to calibrate the procedure concerns size; to evaluate the effect of size we analyzed a paragraph of “A study in scarlet” (Text example), and compared it with a somewhat incoherent speech sample from a patient with schizophrenia (Sample 1). We selected the text example which has evident metaphorical character in stark contrast to Sample 1. In order to establish an even better comparison and analyze further, we examined speech samples generated in response to the question “What activities do people generally do during the course of the day?”, (from Elvevåg et al., 2007) from a healthy participant (Sample 2), and three patients with schizophrenia (Samples 3 to 5). The responses were rated by two human raters for coherence (a score of 1 = very coherence versus 7 = very incoherent) and tangentiality (a score of 1 = very incisively related to question versus 7 = completely unrelated to question).

For all samples we built lexical graphs (see below) and calculated the topological graph parameters (see Section 2.1). We also built topic graphs and calculated the topic and transition entropies (topological graph parameters were not estimated for these graphs because of their small size). Stuttering and repetitions were omitted from speech transcriptions.

[BOX 1]

Text example

“I consider that a man's brain originally is like a little empty attic, and you have to stock it with such furniture as you choose. A fool takes in all the lumber of every sort that he comes across, so that the knowledge which might be useful to him gets crowded out, or at best is jumbled up with a lot of other things so that he has a difficulty in laying his hands upon it. Now the skilful workman is very careful indeed as to what he takes into his brain-attic. He will have nothing but the tools which may help him in doing his work, but of these he has a large assortment, and all in the most perfect order. It is a mistake to think that that little room has elastic walls and can distend to any extent. Depend upon it there comes a time when for every addition of knowledge you forget something that you knew before. It is of the highest importance, therefore, not to have useless facts elbowing out the useful ones.” (p.16; “A study in scarlet” by Arthur Conan Doyle).

Sample 1

“They're destroying too many cattle and oil just to make soap. If we need soap when you can jump into a pool of water, and then when you go to buy your gasoline, my folks always thought they should, get pop but the best thing to get, is motor oil, and, money. May as well go there and, trade in some, pop caps and, tires, and tractors to grup, car garages, so they can pull cars away from wrecks, is what I believed in. So I didn't go there to get no more pop when my folks said it. I just went there to get a ice-cream cone, and some pop, in cans, or we can go over there to get a cigarette. And it was the largest thing you do to get cigarettes 'cause then you could trade off, what you owned, and go for something new, it was sentimental, and that's the only thing I needed was something sentimental, and there wasn't anything else more sentimental than that, except for knick-knacks and most knick-knacks, these cost 30 or 40 dollars to get, a good billfold, or a little stand to put on your desk.” (p. 477; Andreasen, 1986).

Sample 2

“Get up, maybe the alarm clock would wake you up, turn off the alarm clock, or press the snooze bar or something like that, then use the bathroom, brush your teeth, take a shower, maybe shave if you’re a man, then you do your hair, put on clothing, get some breakfast, some people just have coffee or something like that, then go wherever it is you go, school or work, so you might drive yourself, or take the bus or train whatever, to get to where you’re supposed to be for the day, and do whatever it is you’re responsible for doing, working or taking classes, or taking care of your children, or whatever you do during the day, taking breaks during the day for lunch and maybe coffee breaks, or bathroom breaks and at the end of the regular weekday, you go on home by whatever method you came, fixing dinner, or buying something for dinner and eating it, maybe doing some housework or running errands, maybe watching TV or doing something else for recreation, like reading, like a book or a magazine, then get ready for bed, brushing your teeth, putting on pajamas, and getting in the bed.” (Healthy participant; Coherence score: 1, Tangentiality score: 1).

Sample 3

“I'd get up. Usually I take a shower in the morning. Put on clean clothes 'cause I usually slept in the clothes I had on the night before. Eat some kind of breakfast like toast or something. Fix coffee. When I was working, I'd then go to work. Try to get to work by eight. Or go to lunch and eat lunch usually at a restaurant. Then go back to work and work 'til five. Then go home. Then I'd a lot of times go out and have an O'Dooles or a some kind of soft drink, usually a soft drink like um Diet Pepsi or eat a meal like a lot of times I just had salads, but I because I had a hard time with cholesterol so I'd just ate a salad, like lettuce, you know

a side-salad, like lettuce, tomato, onion. Then I'd go home and play on my computer until it was time to go to sleep. I'd usually have the TV on and play the stereo with the usually have the sound up and the TV turned down and on ESPN or whatever sports that was best. And, then I would um usually go to bed about eleven. I'd always take my medicine when I was supposed to would usually take my medicine when I got up in the morning and a lot of times if I had medicine I was supposed to take, I'd take it about lunch time but usually I didn't lately have that then one when I went to bed." (Coherence score: 1, Tangentiality score: 4)

<u>Sample 4</u>

"Ok, a person usually wakes up, at night time you brush your teeth, in the morning you take a shower or bath and you get dressed you feel good, you take a car a cab or a bus or a train to work, and you either go to school or work, and you get something out of it, you get paid, have a good life, and do what suits you.

You come back on the bus or train or the cab, or your car an you go to have a good time you go home, go out to eat, go out with friends, and watch movies, go out to movies and stuff, you stay out of trouble, if you don't stay out of trouble, you go to jail, the worst place to be." (Coherence score: 3.5, Tangentiality score: 4)

<u>Sample 5</u>

"Well, at age forty-seven, I'm waiting around for age forty-eight, to be quite honest.

I'm trying to bear martyrdom of the supreme families, Behovala. Why did you make me forget all those things? I didn't like that. I withhold information. I believe in private property. That's the Bahai faith. At age forty-seven, I'm following Dennis the Menace's mantra.

He followed mantra I want you to be just as good as I am. Stick to your dream and now the LSA can make a responsible decision as to whether you would like another very expensive gift because all labels are one and we have to face that… Is that true or false? Do you want him to suffer again? The end has reached this. That was the medicine I prescribed. Yes, I have many jobs. I've worked with the Wyatt company, I've worked… It's a computer deathbed. It was a good job I worked at a calculator as a businessman. I've worked at

nine to five. In the lower world… Twentieth century AD where I was raised. I was born in eleven-seventy BC. That's an unfair advantage knowing that the Bahai faith comes next. In the early days, they didn't know that. They had to decide if Imagine it is the birth of Buhevaloh, not the birth of Mary and God." (Coherence score: 6, Tangentiality score: 6)

[end BOX 1]

We constructed lexical graphs by linking together words co-occuring in a text at a distance of three words (Ferrer i Cancho, 2005), but removing all function words (e.g., articles, prepositions). We calculated the main graph parameters (clustering coefficient, characteristic path length) of the lexical graphs (see table 2). Notice that all samples show similar measures, but the smallness of the graph precludes us from concluding anything further. As discussed below, future studies using this approach should employ bigger speech samples. Next we focus on the topic graphs.

[PLACE TABLE 2 AROUND HERE]

To obtain the topic graph, we performed a manual labeling procedure (as in the topic graph of “A study in scarlet”), selecting blocks of about one sentence in length. Once the topic graph was built (figure 4), each connected component was assigned a different topic, enabling the calculation of the topic and transition entropies.

[PLACE FIGURE 4 AROUND HERE]

After calculating the entropies, we detected important differences between the patient samples and the controls of comparable length. With reference to table 3, the patient’s discourse (sample 1) results in higher topic and transition entropies than the text example (Holmes). Regarding the responses to the question “What activities do people generally do during the course of the day?” (Samples 2 to 5), it can be seen that the healthy participant’s response (Sample 2) results in lower topic entropy than the patients’ responses. Within the patients’ responses, the one with the lowest coherence (Sample 5) has much higher topic entropy than the others. However, transition entropy was lowest in the healthy participant (Sample 2) and in one of the responses from a patient (Sample 4).

[PLACE TABLE 3 AROUND HERE]

Although preliminary, these results clearly demonstrate the possibility of applying this novel methodological framework to assay the nature of the disorder that is readily apparent in this discourse.

4 Prospects for an automated topic graph construction

One promising direction to automatically obtain the topic graphs is to employ multidimensional semantic spaces. In these, each concept is associated with a vector, a set of concepts is represented as a vector space, and semantic relatedness is gauged as the proximity of the corresponding vectors (e.g., Latent Semantic Analysis (LSA, Deerwester et al., 1990), BEAGLE (Jones et al., 2006)). Semantic spaces are usually built using information on how words co-occur with different frequencies in different contexts. If a large enough corpus (on the order of thousands of documents, each having hundreds of terms) is used, the resulting space can simulate human behavior on a variety of tasks (Landauer and Dumais, 1997; Jones et al., 2006). In order to devise an illustrative automatic procedure, we built an LSA space using 53956 documents and 56108 terms obtained from Wikipedia [Footnote 2], applying standard methods (Landauer et al., 2007). The 390-dimensional semantic space performed comparably well on the TOEFL synonym test (64.65 %, versus the 'gold-standard' of 64.38% (Landauer and Dumais, 1997)) [Footnote 3]. The automatic labeling procedure was as follows: First, we projected each paragraph of text into the semantic space, generating paragraph vectors representing their semantic content. Since each of the terms used to build the word-document matrix can also be represented as a vector in that space, we selected the three terms that were closest to each paragraph and used them as "automatic labels", to build a thematic graph (similar to figure 1 C). We performed the dot product between every word vector and each paragraph vector to determine which word vectors were closest to each paragraph. A pre-selection of words was made by projecting "windows of words" of length 8, and selecting 3 labels for each. Then, when computing dot products for the whole paragraphs, the words were chosen from the previously obtained set of labels, not from the full 56108 terms (figure 5).

[PLACE FIGURE 5 AROUND HERE]

Although the semantic space method produced noisy labels (figure 5), the results are nonetheless encouraging at least when applied to large portions of text. We discuss the potential of this and other methods in section 5.

5 Discussion and future challenges

Communication patterns change across the lifespan, and in illness (e.g., psychopathology and dementia). The convergence of methods from theoretical physics, network theory (Albert and Barabási, 2002), information sciences (Deerwester et al., 1990; Valle-Lisboa and Mizraji, 2007) and cognitive neuropsychiatry (Halligan and David, 2001), presents an opportunity for new frameworks within which to study how humans communicate effectively, and how many pathological processes rob humanity of this most central aspect, namely communicating effectively and meaningfully.

The models and procedures we presented may be valuable modeling tools to assay the underlying structure of discourse disorganization. We have presented a set of tools to analyze text and demonstrate a "proof of concept" of our approach. We believe that if a good topic classification of the speech of patients is achieved, our representations and measures can be valuable tools with which to study schizophrenia. Although promising, the results thus far require subjective judgments to determine the topics a discourse "visits". Also, the graphs and topic classification were generated manually, yet our goal is to devise an automatic method to generate the semantic graph and segment the graph in topics. Ideally our procedure should yield an automatic characterization of incoherent discourse, based on disorganization of the topic graph. Also, the availability of automatically generated large topic graphs would allow a reliable comparison of complex network parameters for normal and pathological samples. We introduced a promising albeit preliminary method based upon LSA (Section 4). Although LSA is a "bag-of-words" approach (as it ignores word-order and syntactic information) it is used in cognitive computational models (Utsumi, 2011) and might have a biological basis (see Mizraji et al., 2009). Alternatively, classification and labeling procedures could rely on neural network models (Dayan and Abott, 2001; Mizraji et al., 2009). Of note, Hoffman (1987) provided an early neural network model to illustrate the putative differences between speech generated from patients with schizophrenia versus those with mania. This model was heuristic by providing a mechanism to understand and visualize specific characteristics of speech, such as perseverative speech versus the seemingly random and rapid associations in mania. Similarly, our work exploits recent developments in network theory and information sciences, as well as the vast computational power available today to construct models of coherent and incoherent discourse. These new technological advances additionally afford the modeling of real data (which is computationally intense), allowing the time-course of discourse to be examined and displaying the results in a visually rich and informative manner. Moreover, these models open up the possibility of building better neural models of the pathophysiology of schizophrenia (Chen, 1994; Talamini et al., 2005; Hoffman and McGlashan, 1997, 1998; Hoffman et al., 1995). Previously, we (Valle-Lisboa et al., 2005) replicated the results of Hoffman, McGlashan and coworkers (Hoffman & McGlashan, 1998, Hoffman et al ,1995) concerning verbal hallucinations, using different models of neural networks (Mizraji, 1989). Our long term goal is to apply neural models to the production of incoherent discourse. If the measurements presented here can be applied generally, and the translation of these procedures to neural models can be achieved, the modeling of language production deviances on a large scale will be possible, and thus provide much needed insights into the neural and cognitive processes underlying speech production in schizophrenia.

References

Albert, R., Barabási, A.L., 2002. Statistical mechanics of complex networks. Rev. Mod. Phy. 74 (1), 47–97.

Andreasen, N.C., 1986. Scale for the assessment of thought, language and communication (TLC). Schizophr. Bull. 12, 474–482.

Andreasen, N.C., Black, D.W., 2005. Introductory textbook of psychiatry. 4th edition. American Psychiatric Association, Washington DC.

Andreasen, N.C., Grove, W.M., 1986. Thought, language and communication in schizophrenia: diagnosis and prognosis. Schizophr. Bull. 12, 348–359.

Cabana, A. 2009. Representación de la estructura del lenguaje escrito mediante grafos y espacios semánticos. [Representation of the structure of written language using graphs and semantic spaces] MSc Thesis. PEDECIBA-Universidad de la República, Uruguay.

Chen, E.Y., 1994. A neural network model of cortical information processing in schizophrenia. I: Interaction between biological and social factors in symptom formation. Can. J. Psychiatry. 39, 362–367.

Collins, A.M., Loftus, E.F., 1975. A spreading-activation theory of semantic processing. Psychol. Rev. 85(6), 407–428.

Collins, A.M., Quillian, M.R., 1969. Retrieval time from semantic memory. J. Verbal Learn. Verbal Behav. 8, 240–247.

Dayan, P., Abbott, L., 2001. Theoretical Neuroscience: Computational and mathematical modeling of neural systems. The MIT Press, Boston.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inform. Sci. 41, 391–407.

DeLisi, L.E., 2001. Speech disorder in schizophrenia: review of the literature and exploration of its relation to uniquely human capacity for language. Schizophr. Bull. 27, 481–496.

Docherty, N.M., 2005. Cognitive impairments and disordered speech in schizophrenia: thought disorder, disorganization, and communication failure perspectives. J. of Abnorm. Psychol. 114, 269–278.

Dorogovtsev, S.N., Mendes, J.F.F., 2001. Language as an evolving word web. Proc. Roy. Soc. B.-Biol. Sci. 268, 2603–2606.

Dumais, S., 1991. Improving the retrieval of information from external sources. Behav. Res. Methods. Instrum. Comput. 23, 229–236.

Elman, J., 1990. Finding structure in time. Cogn. Sci. 14, 179–211.

Elvevåg, B., Foltz, P.W., Rosenstein, M., DeLisi, L.E., 2010. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. J. Neurolinguist. 23, 270–284.

Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. Schizophr. Res. 93, 304–316.

Ferrer i Cancho, R., 2005. The structure of syntactic dependency networks: insights from recent advances in network theory. In: Altmann, G., Levickij, V. , Perebyinis, V. (Eds.) The problems of quantitative linguistics. Chernivtsi: Ruta. pp. 60–75.

Ferrer i Cancho, R., Riordan, O., Bollobás, B., 2005. The consequences of Zipf's law for syntax and symbolic reference. Proc. Roy. Soc. B.-Biol. Sci. 272 (1562), 561–565.

Ferrer i Cancho, R., Solé, R.V., 2001. The small world of human language. Proc. Roy. Soc. B.-Biol. Sci., 268 (1482), 2261–2265.

Foltz, F.W., 2007. Discourse coherence and LSA. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (Eds.) Handbook of Latent Semantic Analysis. Lawrence Erlbaum, New York, pp. 167–184.

Halligan, P.W., David, A.S., 2001. Cognitive neuropsychiatry: towards a scientific psychopathology. Nat. Rev. Neurosci. 2, 209–215.

Hodges, J.R., Salmon, D.P. , Butters, N., 1992. Semantic memory impairment in Alzheimer's disease: failure of access or degraded knowledge? Neuropsychologia, 30, 301–314.

Hoffman, R.E., 1987. Computer simulations of neural information processing and the schizophrenia-mania dichotomy. Arch. Gen. Psychiat. 44, 178–188.

Hoffman, R.E., McGlashan, T.H., 1997. Synaptic elimination, neurodevelopment, and the mechanism of hallucinated "Voices" in Schizophrenia. Am. J. Psychiat. 154, 1683–1689.

Hoffman, R.E., McGlashan, T.H., 1998. Reduced corticocortical connectivity can induce speech perception pathology and hallucinated 'voices'. Schizophr. Res. 30, 137–141.

Hoffman, R.E., Rapaport, J., Ameli, R., McGlashan, T.H., Harcherik, D., Servan-Schreiber, D., 1995. A neural network simulation of hallucinated "voices" and associated speech perception impairments in schizophrenia patients. J. Cogn. Neurosci. 7, 479–497.

Jones, M., Kintsch, W., Mewhort, D., 2006. High-dimensional semantic space accounts of priming. J. Mem. Lang. 55 (4), 534–552.

Jurafsky, D., Martin, J.H., 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall, Upper Saddle River, NJ.

Kintsch, W., 1988. The role of knowledge in discourse comprehension: A construction-integration model. Psychol. Rev. 2, 164–182.

Kuperberg, G., 2010. Language in schizophrenia. Part 1: an Introduction. Lang.Ling. Compass 4 (8), 576–589.

Landauer, T., Dumais, S., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychol. Rev. 104, 211–240.

Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W., (Eds.) 2007. Handbook of Latent Semantic Analysis. Lawrence Erlbaum, New York.

McKenna, P.J., Mortimer, A.M., Hodges, J.R., 1994. Semantic memory and schizophrenia. In David, A.S., Cutting, J.C. (Eds.) The neuropsychology of schizophrenia. Lawrence Erlbaum Associates, Hove. pp. 163–178.

McKenna, P., Oh, T., 2005. Schizophrenic Speech: Making sense of bathroots and ponds that fall in doorways. Cambridge University Press, Cambridge, UK.

Mizraji, E., 1989. Context-dependent associations in linear distributed memories. Bull. Math. Biol., 51, 195–205.

Mizraji, E., Pomi, A., Valle-Lisboa, J.C., 2009. Dynamic searching in the brain. Cogn. Neurodyn. 3, 401–414.

Motter, A.E., de Moura, A.P.S., Lai, Y.C., Dasgupta, P., 2002. Topology of the conceptual network of language. Phys. Rev. E. 65 (6), 065102(R).

Niznikiewicz, M.A., Shenton, M.E., Voglmaier, M., Nestor, P.G., Dickey, C.C., Frumin, M., Seidman, L.J., Allen, C.G., McCarley, R.W., 2002. Semantic dysfunction in women with schizotypal personality disorder. Am. J. Psychiat. 159, 1767–1774.

Palla, G., Derenyi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435 (7043), 814–818.

Pinker, S., 2000. The language instinct. 2nd reimpression. Harper Perennial Classics, New York.

Quillian, M., 1968. Semantic memory. In: Minski, M., (Ed.) Semantic Information Processing, The MIT Press, Cambridge, Massachusetts. pp. 227–270.

Solovay, M.R., Shenton, M.E., Holzman, P.S., 1987. Comparative studies of thought disorder. I. Mania and schizophrenia. Arch. Gen. Psychiat. 44, 13–20.

Steyvers, M., Tenenbaum, J.B., 2005. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. Cogn. Sci. 29 (1), 41–78.

Talamini, L.M., Meeter, M., Elvevåg, B., Murre, J.M., Goldberg, T.E., 2005. Reduced parahippocampal connectivity produces schizophrenia-like memory deficits in simulated neural circuits with reduced parahippocampal connectivity. Arch. Gen. Psychiat. 62, 485–93.

Tamlyn, D., McKenna, P.J., Mortimer, A.M., Lund, C.E., Hammond, S., Baddeley, A.D., 1992. Memory impairment in schizophrenia: its extent, affiliations and neuropsychological character. Psychol. Med. 22, 101–115.

Utsumi, A., 2011. Computational exploration of metaphor comprehension processes using a semantic space model. Cogn. Sci. 35, 251–296.

Valle-Lisboa, J., Mizraji, E., 2007. The uncovering of hidden structures by Latent Semantic Analysis. Inf. Sci. 177, 4122–4147.

Valle-Lisboa, J., Reali, F., Anastasía, H., Mizraji, E., 2005. Elman topology with sigma-pi units: an application to the modeling of verbal hallucinations in schizophrenia. Neural Networks 18, 863–877.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of small-world networks. Nature 393, 440–443.

Footnotes

Footnote 1: The clustering coefficient measures the degree of “socialization” of a node: it is the fraction of a node’s neighbors that are themselves neighbors. A mean clustering coefficient of 0 means that nodes are statistically solitary (in social terms, your friends are not friends of each other), and a coefficient of 1 implies the highest degree of socialization (your friends are all friends of each other). The characteristic path length of a graph is the mean of the lengths of the shortest path between each pair of nodes.

Footnote 2: http://download.wikimedia.org/backup-index.html

Footnote 3: Kindly provided by Prof. T. Landauer. Despite this good score we nonetheless validated all our results using the LSA space available at http://lsa.colorado.edu

Figure legends

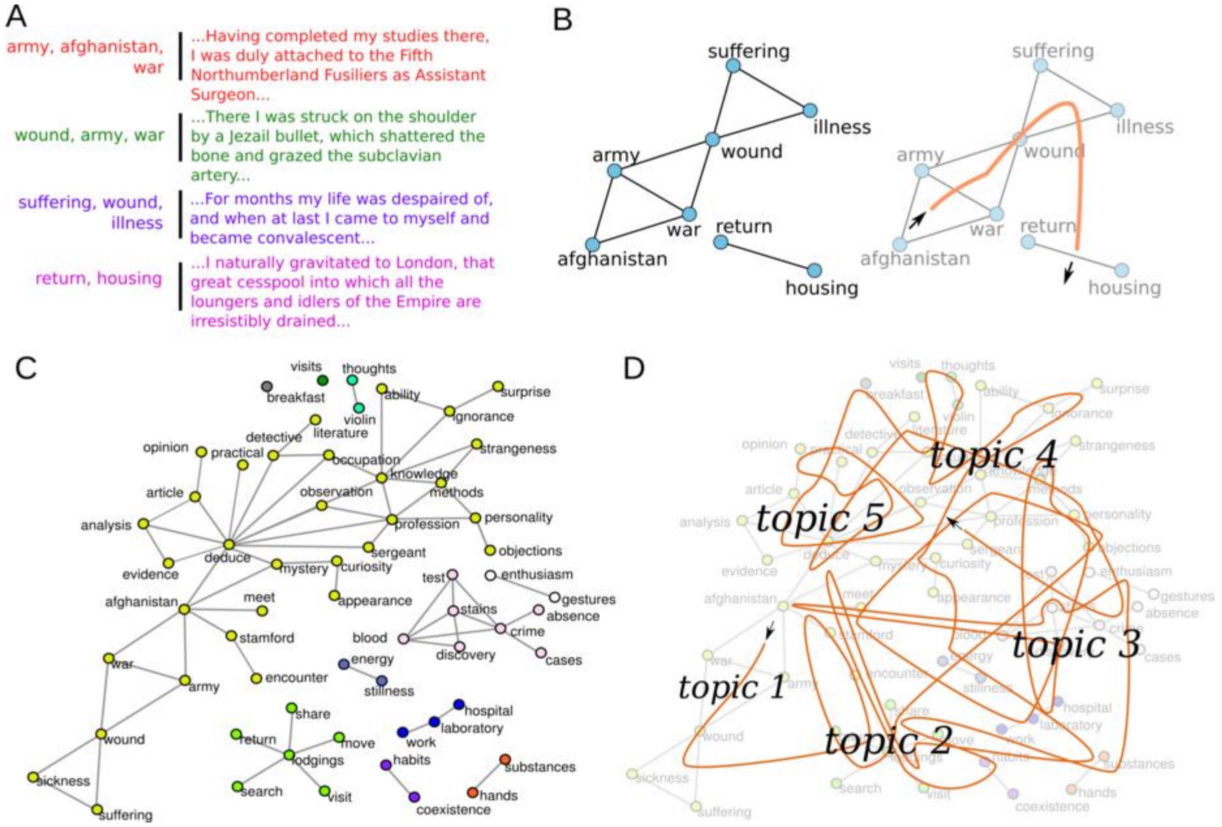
Figure 1: To illustrate our approach, we analyzed the first two chapters of “A study in scarlet” by Arthur Conan Doyle, which features the first appearance of detective Sherlock Holmes (Conan Doyle, 2005). The first paragraphs and an illustration of the labeling process are shown:. A) Four thematic blocks with labels are delimited. B) The resulting graph after assigning a node to each label, and linking labels that occur together in a block assignment, called topic graph. C) The resulting topic graph of the two chapters. D) The discourse trajectory is drawn over the topic graph.

Figure 2: A) Schematic diagram of the shuffling procedure. Two “cutting points” are randomly assigned in the text, and then the remaining three portions of text are randomly permuted. B) Visualization of the effectiveness of the shuffling procedure on the topic assignment of the first two chapters of “A study in scarlet”, based on the 5 topics identified in the thematic graph (figure 1 D). Note how the mixture develops as the shuffling is iterated.

Figure 3: Topic and transition entropy increase when the text of the first two paragraphs of “A study in scarlet” is shuffled. A) The topic entropies, calculated using equation (1). B) Transition entropy of each topic, according to equation (3).

Figure 4: Topic graphs obtained from manual label assignment for A) Sherlock Holmes' “speech” , B) a patient with schizophrenia (Sample 1), C) a healthy participant (Sample 2), D, E and F) patients with schizophrenia (Samples 3 to 5). Discourse trajectories are shown as lines over the graphs.

Figure 5: Topic graph obtained by applying the automatic procedure to the first two chapters of “A study in scarlet”. Although the resulting labeling is not optimal, this automatically generated graph has a central component and several topics that can readily be detected, enabling the computation of entropies. Topic entropy was 5.13 and transition entropy was 4.38, values that compare well with entropies generated via the manually labeled topic graph (4.84 and 4.76, respectively).



army, afghanistan, war | ...Having completed my studies there, I was duly attached to the Fifth Northumberland Fusiliers as Assistant Surgeon...

wound, army, war | ...There I was struck on the shoulder by a Jezail bullet, which shattered the bone and grazed the subclavian artery...

suffering, wound, illness | ...For months my life was despaired of, and when at last I came to myself and became convalescent...

return, housing | ...I naturally gravitated to London, that great cesspool into which all the loungers and idlers of the Empire are irresistibly drained...

FIGURE 1

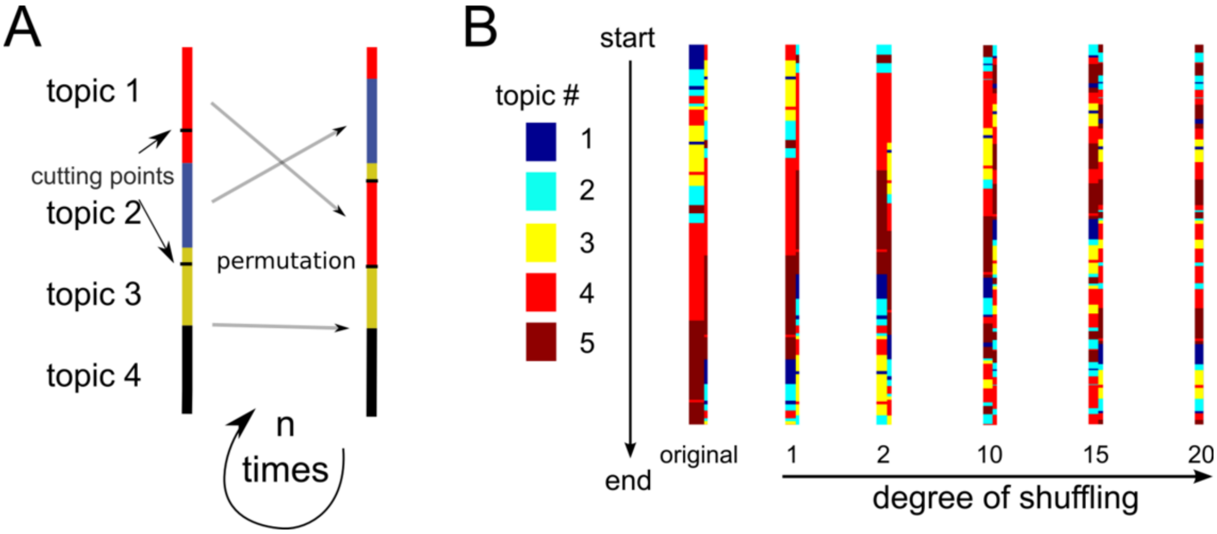

A
topic 1
cutting points
topic 2
topic 3
topic 4
permutation
n times
B
start
end
topic #
1
2
3
4
5
original
1
2
10
15
20
degree of shuffling

FIGURE 2

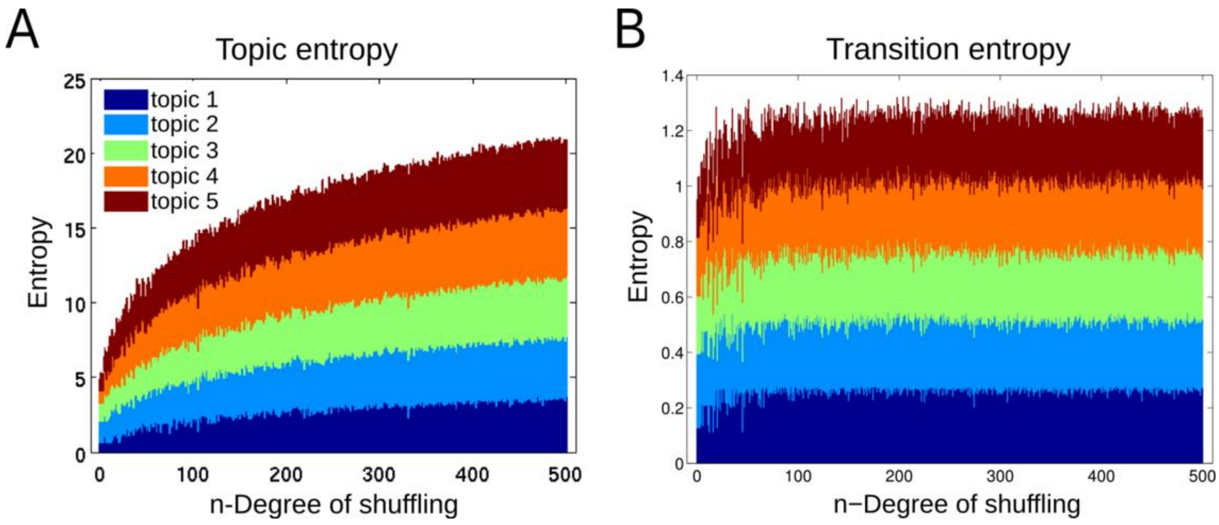

A
Topic entropy
topic 1
topic 2
topic 3
topic 4
topic 5
Entropy
0
5
10
15
20
25
0
100
200
300
400
500
n-Degree of shuffling
B
Transition entropy
Entropy
0
0.2
0.4
0.6
0.8
1
1.2
1.4
0
100
200
300
400
500
n−Degree of shuffling

FIGURE 3

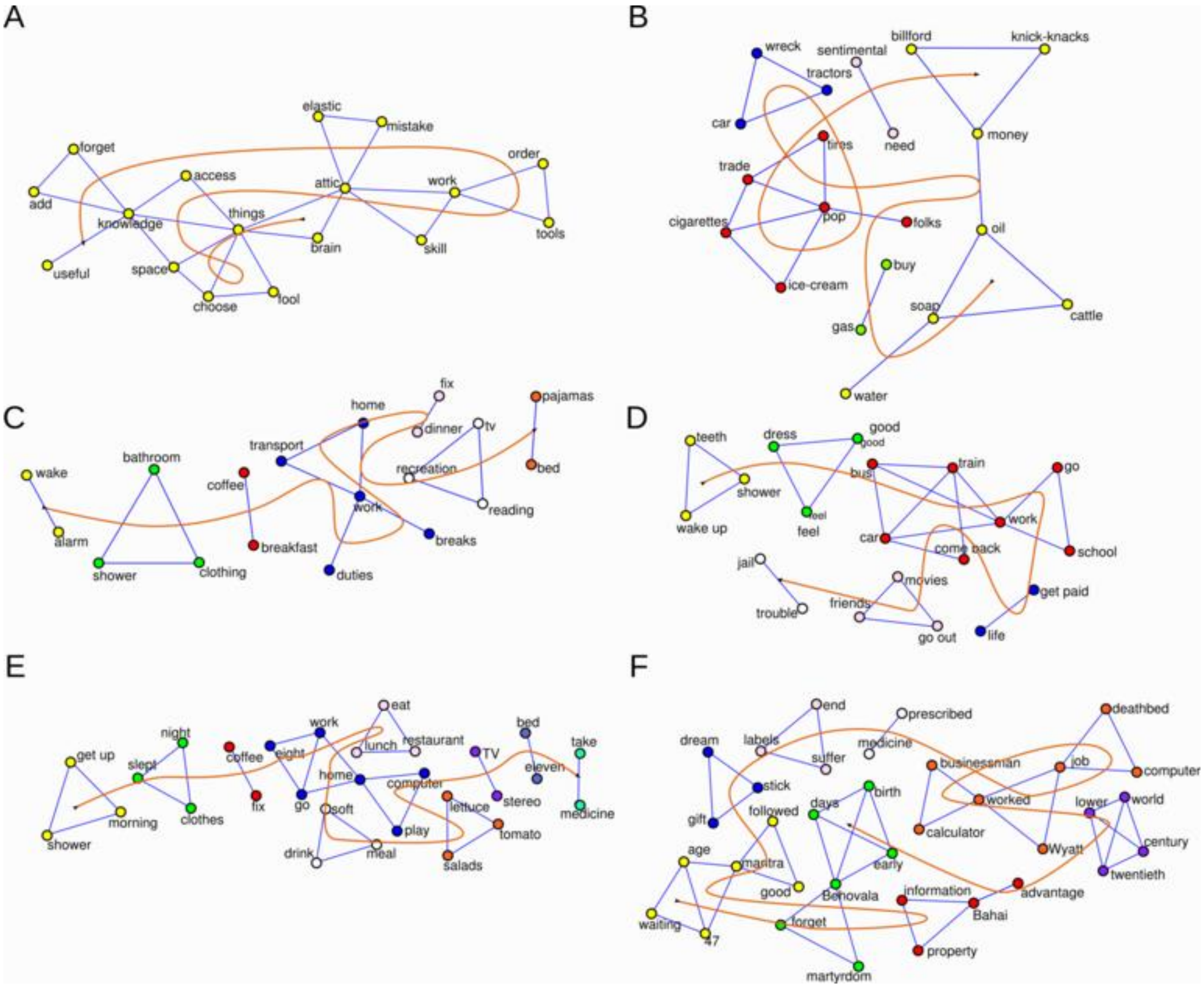

A
forget
add
knowledge
useful
access
things
space
choose
fool
elastic
mistake
attic
brain
work
skill
order
tools
B
wreck
sentimental
billford
knick-knacks
tractors
car
tires
need
money
trade
pop
folks
cigarettes
oil
buy
ice-cream
soap
cattle
gas
water
C
fix
pajamas
home
dinner
tv
transport
recreation
bed
wake
bathroom
coffee
work
reading
alarm
breakfast
breaks
shower
clothing
duties
D
teeth
dress
good
good
shower
bus
train
go
wake up
feel
feel
work
car
come back
school
jail
movies
get paid
friends
trouble
go out
life
E
eat
work
bed
take
night
restaurant
get up
lunch
TV
slept
coffee
eight
eleven
home
computer
fix
go
soft
stereo
lettuce
medicine
morning
clothes
play
tomato
shower
drink
meal
salads
F
end
prescribed
deathbed
dream
labels
medicine
suffer
businessman
job
computer
stick
birth
days
worked
lower
world
gift
followed
calculator
Wyatt
century
age
early
mantra
twentieth
good
Behovala
information
advantage
waiting
forget
Bahai
47
property
martyrdom

FIGURE 4

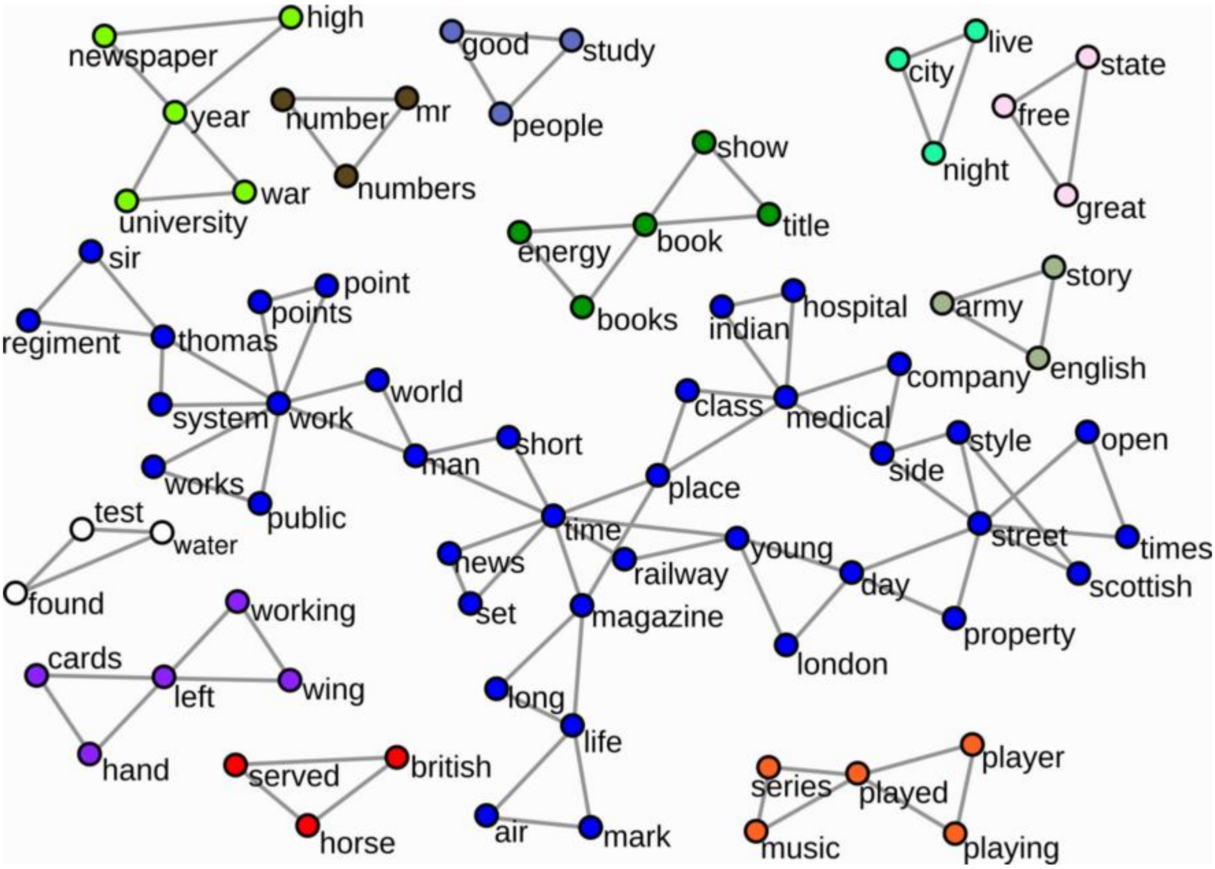

high
newspaper
year
war
university
number
mr
numbers
good
study
people
show
book
title
energy
books
live
city
night
state
free
great
sir
regiment
thomas
point
points
system
work
world
works
public
man
short
time
news
set
railway
magazine
young
day
london
indian
hospital
class
medical
company
place
side
style
open
street
times
scottish
property
story
army
english
test
water
found
working
cards
left
wing
hand
served
british
horse
long
life
air
mark
series
played
player
music
playing

FIGURE 5

Tables

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Afghanistan	return	crime	personality	mystery
war	search	cases	strangeness	deduce
wound	lodgings	blood	methods	evidence
sickness	move	laboratory	occupation	analysis
suffering	coexistence	substances	knowledge	detective

Table 1: Representative topic labels used to categorize each sentence of the text.

Network	Clustering Coefficient	Path Length	Nº of nodes	Links
Text example (Holmes)	0.55	3.63	49	109
Sample 1 (Patient)	0.51	4.04	45	95
Sample 2 (Control)	0.54	4.39	56	121
Sample 3 (Patient)	0.50	3.02	49	124
Sample 4 (Patient)	0.52	3.34	34	72
Sample 5 (Patient)	0.51	4.41	61	128

Table 2: Characteristics of lexical graphs obtained from the samples.

	Nº of topics	**Sum of Topic entropy**	**Sum of transition entropy**	**Mean topic entropy**	**Mean Transition entropy**
Text (Holmes)	1	0	0	0	0
Sample 1 (Patient)	4	1.94	1.79	0.38	0.35
Sample 2 (Control)	7	0.30	0.69	0.043	0.099
Sample 3 (Patient)	10	1.08	1.10	0.11	0.11
Sample 4 (Patient)	6	0.67	0.69	0.11	0.12
Sample 5 (Patient)	8	2.05	1.39	0.26	0.17

Table 3: Comparison of topic and transition entropy for the samples. In each case the sum of the entropy for each topic is computed. Normalizing by the number of topics we obtain the mean topic and transition entropy, displayed in columns 5 and 6 respectively. Entropies for the text example are zero because only one topic was detected.

Role of Funding Source

EM and JCVL were supported by PEDECIBA and CSIC-UDELAR. AC was supported by PEDECIBA. BE was supported by the Northern Norwegian Regional Health Authority (Helse Nord RHF). None of the funding agencies (PEDECIBA, CSIC or Helse Nord RHF) had any further role in study design, data analysis and interpretation, or in the writing of the report and decision to submit for publication.

Contributors

EM and JVL conceived the original ideas and theoretical framework. AC designed the procedures to create and analyze the lexical and topic graphs, corrected and improved the LSA-Wikipedia previously created by JVL and wrote all the programs for text and network analysis. BE motivated the clinical application of the theoretical framework and tools. All authors discussed the methods and results and contributed to the writing of the manuscript. All authors have read and approved the final manuscript.

Conflict of Interest

None of the authors have any potential conflicts of interest or biomedical financial interests.

Acknowledgements

We thank Dr. Andrés Pomi for valuable discussions and encouragement.