Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Long range dependence in texts: A method for quantifying coherence of text

Elham Najafi*, Amir H. Darooneh

Department of Physics, University of Zanjan, P.O.Box 45196-313, Zanjan, Iran

ARTICLE INFO

Article history:
Received 12 March 2017
Revised 21 May 2017
Accepted 25 June 2017
Available online 27 June 2017

Keywords:
Global coherence
Importance time series
Scaling exponent

ABSTRACT

This paper discusses a major issue in computational linguistics; the automatic calculation of text coherence. Heretofore, only few methods have been proposed to automatically detect local coherence of texts. All of these methods need a lot of pre-processing tasks and computational efforts. Here we suggest a simple method to evaluate the coherence globally. First, we use a word ranking method to assign an importance value to each word-type in a text, then the importance time series associated with text is constructed. In the next step, Detrended Fluctuation Analysis(DFA) which is used for detecting inherent correlations in time series, is applied to texts importance time series. We found that the importance time series exhibits a bi-scale behavior; it is long-range correlated at large distances, while short-range correlations are observed in small distances. We also observed that for a shuffled text the scaling exponent decreases. This decrease becomes more and more significant when we reshuffle the chapters, paragraphs, sentences and words respectively. This fact leads us to consider the scaling exponent of text time series (or briefly STT) as a measure for quantifying the global coherence. We demonstrate our claim by carrying out an experiment on three sample texts and comparing our method by some entity grid based models.

1. Introduction

Much of the human cognition can be interpreted as a mental process for changing the sensory perceptions into coherent patterns of concepts. This coherence demonstrates itself in the ordering of words when one tries to communicate his ideas by writing or speaking. Quantifying the coherence in texts may help us to explore some hidden aspects of cognitive processes [1–5]. Automatic coherence evaluation is also an interesting subject in artificial intelligence, linguistics, data mining and some other disciplines.

Coherence is a quality of written or spoken texts that makes them easy to read and understand [15,16]. A coherent text obeys a particular logical order and is easy to comprehend as a unit, instead of a bunch of messy sentences. Text coherence occurs in two levels; local and global. Local coherence is representative of similarity among adjacent parts of a text, for example, adjacent paragraphs or sentences. On the other hand, global coherence is indicator of the connection between all segments of a text as a whole [17].

Text is an ordered sequence of words that can be considered as an one dimensional discrete space. The meaning of text regulates the distribution of words throughout this space. All word types have self similar distributions across the text but with different fractal features. An importance value can be assigned to each word type by considering its fractal pattern [6]. We associate a time series to every text by substituting the words with their corresponding importance values while retaining their order. The importance time series exhibits the long range dependence for meaningful texts. We assert in this work that the long range correlation in the importance time series is related to the global coherence of the text.

The detrended fluctuation analysis (DFA) is the most powerful method for exploring the long-range correlations in non-stationary time series [8–10]. We use several variants of this method for analyzing three sample texts. We find that importance time series of texts displays bi-scale behavior; There is short range correlation between words at small distances while for larger distances the long range correlation is observable. The text shuffling causes the long range correlation to disappear. We shuffle texts in various levels: Chapters, Paragraphs, Sentences and Words. In each level, text is comprises of some sections (E.g. in paragraph level, the number of paragraphs in a text indicates the number of sections.) Each section of a text has a position inside the text. Lets indicate this position with label i. In shuffling process the position of section i randomly changes to i'. In a way that no sections can be found in the same position. Shuffling the chapters, paragraphs, sentences

* corresponding author.
E-mail addresses: e.najafi@znu.ac.ir, elhamnajafi.272128@gmail.com (E. Najafi), darooneh@znu.ac.ir (A. H. Darooneh).

http://dx.doi.org/10.1016/j.knosys.2017.06.032
0950-7051/© 2017 Elsevier B.V. All rights reserved.

and words result in more and more decrease in the long range correlation respectively. This fact persuades us to measure the text coherence by analyzing the auto correlation function. It is worth noting, there are other methods which determine importance of the words in a given text according to their spatial distributions. These methods can be also used to construct importance time series. We show here that they lead to the similar results for global coherence.

Heretofore, some methods have been suggested for evaluation of text (local) coherence. Two main techniques are latent semantic analysis (LSA) and the entity grid method. Foltz et al. in 1998 [18] introduced the latent semantic analysis approach for computing the local coherence of a text [12]. LSA is a vector-space model for measuring semantic similarity of a set of documents or parts of a document. Firstly, a term-passage matrix is constructed, that its rows indicate individual terms (word types or phrases), and the columns show different passages. Each matrix element shows the value that is assigned to a term in a particular passage. The prevalent choice is the term frequency in the given passage. The similarity between two passages is defined as cosine of the angle between their columnar vectors. The local coherence of a text is average of all similarity values. In practice, in order to reduce the number of rows with preserving the similarity structure among columns, singular value decomposition (SVD) is used. Another method also exists that is slightly different from LSA in definition of the term-passage matrix but it uses the same computational technique for evaluating the local coherence [19].

Barzilay and Lapata [11] have proposed another method for detecting the local coherence. The main assumption of this method is that there are some regularities in entity distribution of locally coherent texts. The entities are classified by their grammatical roles as subject, object or neither. The entity grid can be constructed as a matrix which its elements are the grammatical role of a word in a sentence. The rows represent the different sentences and the columns are representative of the word types. A word may be subject in a particular sentence and object in the next, then the subject to object transition has occurred for this word. Such transition is enumerated over all words. There are sixteen transition possibilities, probability for all these transitions can be computed and put in a vector with sixteen elements. This is a feature vector that is associated with each text and is related to its coherence. It is possible to rank the texts in a corpus according to their feature vectors by using one of the supervised learning models. Some other investigations have also been developed based on entity grid model to improve the ranking efficiency [13,14,20–22].

LSA is weakly sensitive to shuffling process and practically is not applicable to the case of global coherence [12]. The entity grid method only ranks the texts within a corpus and cannot measure the coherence for an individual text. Furthermore, several preprocessing tasks are needed for both methods to increase their performance.

In contrast to considerable investigations on automatic local coherence, most of the works on global coherence are manual methods [24]. Here, for the first time we propose an automatic method for calculating the global coherence of texts which needs no preprocessing and is remarkably sensitive to text shuffling. In this method scaling exponent of text time series (STT) is considered as text coherence and the results are compared with some coherence models.

In the following section, we review the fractality method which is used to assign an importance value to each word-type in a text. Then we explain about constructing importance times eies from texts. Furthermore we explain Detrended Fluctuation Analysis (DFA) model for calculating scaling exponent of the text time series. In Result and Discussion section, we apply our proposed method to calculate coherence of three sample texts and compare the results with some entity grid based methods. In the final section we represent a conclusion.

2. Methods

Text is a sequence of words that are ordered to express writer's thoughts and senses so that it is understood by readers. The pattern of distribution of each word across the text, is mostly controlled by the meaning. The meaning is a unit entity which is unfolded through chapters, paragraphs, sentences and words. Therefore, coherence as a measure for text's unity is determined by studying the word types distributions in text. The spatial distribution of a word type shows how it is important in conveying the text's meaning. In the following, we describe computation of the importance value for a word and constructing the importance time series. We also discuss analyzing the time series by DFA method and calculating the scaling exponent. The scaling exponent is a characteristic measure for quantifying the long range correlation and then the global coherence of text.

2.1. Fractality: an importance measure

A text is a collection of well-positioned words. In a text, words are ordered with a particular arrangement. This particular arrangement arises for two reasons. First, grammatical rules determine where words should be placed within a sentence. These rules indicate the positions of verbs, nouns, adverbs, and other parts of speech and make short range correlations between the sequences of words in a sentence. Secondly, a text gets meaning from the specific order which the words are distributed throughout. This ordering is long-range (i.e. acts across the whole range of the text) and is called semantic ordering. These rules impose different word types to have different importance values in a text.

Najafi and Darooneh used the concept of fractals to introduce fractality of a word as an index for its importance [6]. Fractal is a mathematical object or a set of points in space which has a repeating pattern in every scales. Fractal dimension represents how details of a fractal pattern varies by changing scale [23]. A text is considered as an one dimensional array of words. The places of occurrence of any word type forms a self similar set where it is embedded in this one dimensional discrete space. The fractal dimension measures the self similarity of such sets. There are two kinds of words in text: words which are related to the subject of the text; namely the important words, and all others that are irrelevant to it. For instance for a text in cosmology, words like *universe, space, big-bang, inflation, etc.*, are important words. The other words which are found in every texts such as: *the, of, is, it, have, happening, etc.*, are common or irrelevant words. By calculating fractal dimension of words in a text, the important words of a text could be distinguished. The fractal dimension of words in this one-dimensional space is a number between 0 and 1. The important words have dimensions that are significantly different from one, while the less important words are distributed uniformly and their dimensions are close to one [6]. Box counting is a practical method for computing the fractal dimension of words. To perform this calculation, the text should be divided into boxes of size s. It means that any box contains s successive words. The number of such boxes that contains at least one occurrence of a given word w, i.e. the number of filled boxes, is $N_b(w, s)$ and the self-similarity property is expressed as,

$$N_b(w,s) \sim s^{-D_w}. \tag{1}$$

where D_w stands for the fractal dimension of word w. It is obtained by finding the slope of log-log plot of $N_b(w, s)$ versus s for large enough s.

The shuffling process cannot change the pattern of the words that are uniformly distributed throughout the text, although meaning of the text gets lost. Therefore, such words have less importance in the semantic structure of the text. The distributions of the more important words change remarkably in the shuffled text. Accordingly, for an important word, the number of filled boxes in an original text differs from the case of the shuffled text. Difference between the number of filled boxes for a given word in the original and shuffled text could be considered as a measure of word importance [6]. To measure these differences the *degree of fractality* is defined as:

$$d_f(w) = \sum_s \log(\frac{N_b^{sh.}(w,s)}{N_b(w,s)}) \tag{2}$$

where $d_f(w)$ is the degree of fractality (or simply fractality) and $N_b^{sh.}(w,s)$ is the number of filled boxes with size s for the particular word w, in the shuffled text. To take into account the frequency of each word in a text, the fractality could be multiplied by $log(M)$, where M is word frequency, as is explained in the reference [6].

To calculate the number of filled boxes in a shuffled text, we have to perform one shuffling process for each particular word. It means that we need to perform a large number of shuffling processes (equal to the number of word-types in a text) to rank the words due to their importance. To overcome this difficulty we use our conjecture about the number of filled boxes in a shuffled text [6]:

$$N_b^{sh.}(s,\omega) = \frac{M}{1+(\frac{M-1}{N-1})(s-1)} \tag{3}$$

where M is frequency of the word ω. This equation shows good conformity with the number of filled boxes in a shuffled text [6]. Higher value of degree of fractality means that the distribution pattern of the word is more different from the uniform distribution; So the word is more important.

2.2. Importance time series of texts

In text mining, several ways exist to convert a text to time series which are used for certain purposes. Here we describe how to build the importance time series from a given text. It is believed that many informative aspects of the text are included in these time series.

As is mentioned earlier, with every text, we can associate a one-dimensional array. The successive words in text are placed in consecutive positions in the array. This array is directed from beginning word of the text to the last one, in accordance to the time direction that the text is written. The importance time series associated with the text is simply constructed by replacing each word with its importance value in the aforesaid array. All of the word ranking methods are admissible for assigning importance values to the set of vocabulary words. So, we can build several time series that are constructed from a given text. In the following we explain how DFA can be used for detection of the long range correlation in a time series and determination of its scaling exponent.

2.3. Detrended fluctuation analysis

Time series with long range correlation have the power law decaying auto-correlation function. A scaling exponent characterizes such behavior. Calculation of this exponent is the main purpose of all methods for analysis of the long range correlations in time series. In most of the data there are some trends that cause the time series to be non-stationary and also the non-intrinsic long range correlation to appear. These trends have unpleasant influence on the value of scaling exponent. Hence trends should be eliminated from time series to obtain a more reliable value for the scaling exponent. DFA is the most powerful method for detecting the long range correlations in non-stationary time series [9,10]. The advantage of DFA is elimination of the trends when the scaling exponent is computed. We briefly explain the method here.

The DFA procedure consists of four steps:

1. For time series $\{x_1, x_2, \ldots, x_N\}$, the global profile, $X(i)$, is defined as;

$$X(i) = \sum_{k=1}^{i} x_k - <x>, \tag{4}$$

where $<x>$ is the average of data xs.

2. The profile is divided into $N_s = \lfloor N/s \rfloor$ non-overlapping segments of size s. Since the time series length may not be a multiple of segment size s, a small number of data at the end of time series will remain. To prevent loosing that small part of data, the division of the profile is repeated from the other end of the time series. So, in total $2N_s$ segments are obtained.
3. In each segment, the local trend is calculated by a least-square fitting of the data. Then, the trend is subtracted from the profile and detrended profile for segments of size s is calculated:

$$X_s(i) = X(i) - p_n(i), \tag{5}$$

Where $p_n(i)$ is the fitting polynomial in the nth segment.

4. The variance of the detrended profile in each segment n is calculated by averaging over all data points i:

$$F_s^2(n) = < X_s^2(i) > = \frac{1}{s}\sum_{i=1}^{s} X_s^2[(n-1)s+i], \tag{6}$$

Finally, the variance is averaged over all $2N_s$ segments. The square root of this average is the fluctuation function:

$$F(s) = [\frac{1}{2N_s}\sum_{n=1}^{2N_s} F_s^2(n)]^{1/2}. \tag{7}$$

Different orders of fitting polynomials lead to different detrending orders. When linear polynomials are used, the fluctuation analysis is called DFA1. DFA2 removes second order trends of the profile. Higher order polynomials can also be used in the fitting procedure.

The procedure for DMA [28,29] method is same with only one difference: $p_n(i)$ is replaced by $y_n(i)$ which is the moving average of data in the time series. The moving average function is calculated by averaging over $n/2$ past and $n/2$ future points in each sliding window with size n.

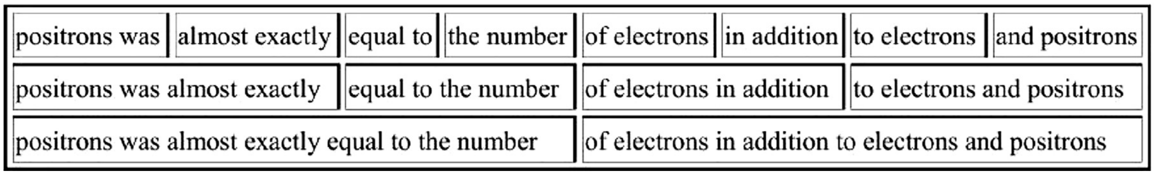
By increasing s, F(s) increases in a power-law form:

$$F(s) \propto s^{\alpha} \tag{8}$$

α is the scaling exponent of time series. The Value of scaling exponent indicates the range of correlation in a time series. $\alpha = 0.5$ indicates a short-range correlation like a random walk. The case of $0.5 < \alpha < 1$ corresponds to a persistent long-range power-law correlation; it means that each large value in a time series is more likely to be followed by another large value. In contrast $0 < \alpha < 0.5$ indicates an anti-persistent power-law correlation such that each large value in the time series is more likely to be followed by a small value. We assert that α exponent of text time series could be an index for coherence of a text.

3. Result and discussion

To examine our method, we peruse three sample books, *On the Origin of Species by Means of Natural Selection* by *Charles Darwin* [30] which is about evolution of populations through a process of natural selection, *Relativity: The Special and General Theory*



positrons was | almost exactly | equal to | the number | of electrons | in addition | to electrons | and positrons

positrons was almost exactly | equal to the number | of electrons in addition | to electrons and positrons

positrons was almost exactly equal to the number | of electrons in addition to electrons and positrons

Fig. 1. A sample text chosen from *The First Three Minutes* and its divisions into boxes. Box size for first row is equal to 2 and for the second and third rows are 4 and 8 respectively.

Table 1
Some properties of our sample texts. N is the number of words in the text, N_v is the number of vocabularies (word types), N_c is the number of chapters, N_p number of paragraphs, and N_s is the number of sentences in each text.

Book's name	N	N_v	N_c	N_p	N_s
On The Origin of Species	191740	8842	15	946	4793
Relativity	30089	2762	36	270	832
The First Three Minutes	48476	4039	9	258	1458

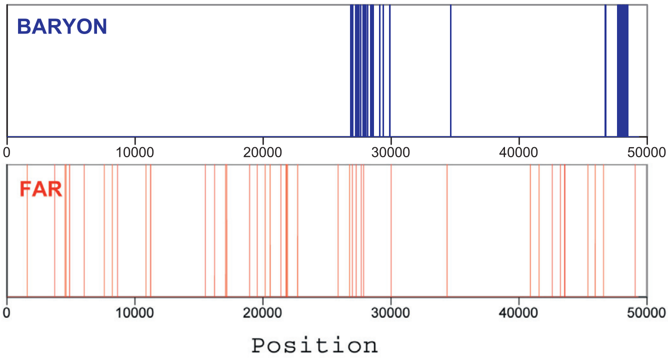


Fig. 2. Spatial distribution of two words, *BARYON* and *FAR*, in the book, *The first Three Minutes*. According to subject of the book, *BARYON* is an important and the *FAR* is an irrelevant word, both of them have the same frequency equal to 45. *FAR* is distributed uniformly in the text, but *BARYON* is clustered.

by *Albert Einstein* [31] about general and special relativity, and *"The First Three Minutes"* by *Steven Weinberg* [32] which is about creation of matter based on the standard model in the big bang scenario. Some useful properties of the books can be found in Table 1.

In first step, we should calculate the importance values for all word types in our sample texts. To calculate the fractality values, text should be divided into non-overlapping boxes of size s and the number of filled boxes (N_b) should be counted. Fig. 1 displays dividing part of *"The First Three Minutes"* text into boxes. The first row of the figure shows boxes with size 2($s = 2$). For the second row $s = 4$ and for the third row $s = 8$ is chosen. As an example consider the word *electron*. It has appeared in 2 boxes in first raw. So, the number of filled boxes with size $s = 2$ for this word equals to 2. In the same way, we can see $N_b = 2$ for box sizes $s = 4$ and finally $N_b = 1$ for $s = 8$. The number of filled boxes is used to indicate fractal dimension of words.

The distribution of an important word throughout a text is considerably different from distribution of an irrelevant word. Fig. 2 shows the spatial distribution of two words, *BARYON* and *FAR*, in the book, *The first Three Minutes*. According to the subject of the book, *BARYON* is an important word and the *FAR* is an irrelevant word. As is seen from the Fig. 2, *FAR* is distributed uniformly in the text, but *BARYON* is clustered.

Because of the uniform distribution of irrelevant words in a text, shuffling process do not change the distribution of them significantly. But the distribution of important words alter considerably when the text undergoes a shuffling process. Fig. 3 shows the result of box counting for distribution of *BARYON* in the original and shuffled text. *BARYON* is an important word in the book, *The First Three Minutes*. As is seen from this figure, there is a considerable difference between box-counting of this word in the original and shuffled text. The plotted dashed line is our conjecture about the number of filled boxes in a shuffled text. If we shuffle a text N times we would see that the result of box counting for shuffled text is very close to our conjecture. It seems that this conjecture is a good approximation of number of filled boxes in a shuffled text. As is seen from the figure there is a good conformity between the conjecture and the data.

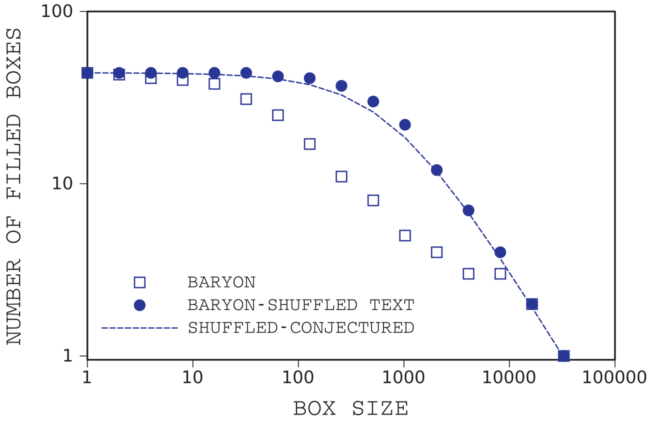


Fig. 3. Results of box counting for distribution of *BARYON* in the original and shuffled text. *BARYON* is an important word in the book, *The First Three Minutes*. There is a considerable difference between box-counting of this word in the original and shuffled text. The dashed line is our conjecture about the number of filled boxes in a shuffled text.

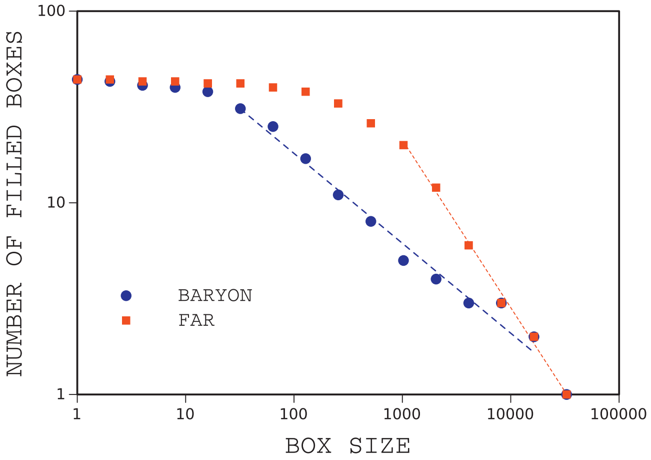


Fig. 4. The number of filled boxed versus box size for *BARYON* and *FAR* from *The First Three Minutes*.

Fig. 4 shows the diagram of filled boxes versus box size for *BARYON* and *FAR* from the book *The First Three Minutes*. *BARYON* is an important word in the text and *FAR* is irrelevant. So, we expect to see different distribution patterns and therefore different fractal dimensions for these two words. The slope of diagram of the number of filled boxes versus box size indicates fractal dimension of corresponding words. As is seen from this figure, the fractal di-

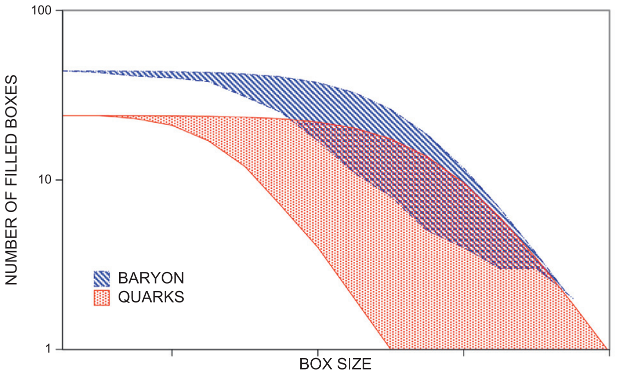


Fig. 5. The area between diagram of filled boxes in original and shuffled version of *The First Three Minutes*. The blue-dashed area corresponds to *BARYON* and red-dashed area corresponds to *Quark*. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
List of the twenty top-ranked words according to fractality (left) and the first twenty common words (right) from *The first three minutes*. Words with high degree of fractality are important words according to subject of the book and common words have low fractality values.

Words	Frequency	Fractality	Words	Frequency	Fractality
quarks	24	17.808	the	4159	2.3974
ions	14	16.8488	of	2447	2.4961
diagram	7	15.295	to	1154	2.61037
rotating	7	13.7272	a	1078	2.5462
quark	10	13.4022	in	1060	2.5577
hydroxyl	7	12.8799	and	945	2.5635
nebulae	14	12.4461	is	898	2.4926
kmsec	10	12.239	that	731	2.6757
antenna	35	12.227	universe	502	2.2157
stone	6	12.1493	this	434	2.5273
slab	4	12.0712	it	422	2.5563
luminosity	24	11.508	as	418	2.9524
cyanogen	10	11.3873	be	395	2.6094
angstroms	4	11.378	at	391	2.5323
circumference	4	11.2603	by	362	2.8173
conservation	23	10.9895	we	343	2.9049
mc	8	10.4745	for	325	2.1734
planck	18	10.4436	are	299	2.8623
candidate	2	10.3972	was	293	2.1275
disc	6	10.37	with	289	2.7564

mension of these two words are considerably different from each other. The fractal dimension is about 0.5 for BARYON and is close to 0.9 for FAR. So, the fractal dimension of important words is different from irrelevant ones. The fractal dimension of words can be used as a measure of word importance and can be used to rank words of a text due to their importance values.

For practical reasons, the fractal dimension could not be calculated precisely. So, we use the difference between distribution of words in an original text and its shuffled version, to introduce an importance measure, which is called Fractality [6]. The reason for this choice is that shuffling process significantly change the distribution of important words. But, because of uniform distribution, the distribution of irrelevant words do not change considerably. In Fig. 5 the blue shaded parts of the plot shows the difference between number of filled boxes for *BARYON* in the original text and its shuffled version.

The more the difference between these two diagrams, the more important that word is. As is seen in Fig. 5 the word *Quarks* is more important than *BARYON* in the book *"The First Three Minutes"*. In Table 2 we have listed twenty most important words based on *fractality* for one of our sample books; *The first three minutes*. According to this table, the top-ranked words are important according to subject of the text. Besides, the list of most frequent words are available in Table 2. As is seen, this frequent words have low fractality values.

The next step is constructing the important time series. For this purpose, an array which is directed from beginning word of text to the last one is considered. The importance time series of the text is simply constructed by replacing each word with its importance value. In Fig. 6, a schematic of constructing fractality time series from a sample text chosen from *The First Three Minutes* is displayed. As is seen in this figure, the fractality values are normalized in order to be a number between 0 and 1. We do the normalization because, the fractality values of word-types in a text depend on the text size. It means that the maximum of fractality value increases when the text is longer. So, we normalized the fractality values to make the time series of different texts comparable. It is worth noting that the correlation exponents obtained from DFA method is not influenced by normalizing the time series. For constructing the time series of Fig. 6, each word is replaced by normalized value of fractality multiplied log(M).

In Fig. 7 we show the fractality time series of the first 500 words of our three sample books.

With help of DFA method, we detect the correlations in these time series. In Fig. 8a the result of DFA1 for fractality time series is plotted for the book *"On the Origin of species"* by blue circles. The scaling exponent α, is obtained by power regression of this diagram. As is seen in the figure, this time series should be described by two different α exponents. One α exponent for small distances between words; around 150 and 200 for our sample books, and another exponent for distances larger than this value. In other words, the time series shows bi-fractal behavior in small and large scales. By power regression the scaling exponent for small s is $\alpha = 0.5$ and for large s, $\alpha = 0.73$. So, in small scales the time series is short-range correlated and for large scales it exhibits long-range correlation.

We used different orders of DFA and also DMA method for analyzing all of our sample books and bi-scale behavior of texts was observed. The result of DFA for *Relativity* is shown in Fig. 8b and the result of DFA for *The First Three Minutes* is shown in Fig. 8c. According to these figures, the fractality time series of these texts are again bi-scale and also, long-range correlated.

If a text undergoes any shuffling process, text coherence decreases. Besides, by shuffling a text, the long-range correlation of its time series is expected to decrease. So, there should be a relation between text coherence and scaling exponent of its time series. We assert that scaling exponent of text time series may be an indicator of text coherence.

To study this point more clearly, we shuffle our sample texts in different levels (i.e., chapters paragraphs, sentences and words), then compute the fractality of words in the shuffled texts and construct time series from these shuffled texts. Finally, we calculate scaling exponent for these shuffled versions of the sample texts. With shuffling the texts, level by level, text coherence decreases gradually and we also expect to observe decreasing of scaling exponent in every step. To check the accuracy of our expectation, we carry out the shuffling process in different levels, including chapters, paragraphs, sentences and words for our sample books and consider the changes in α exponent. Fig. 8a shows the result of DFA1 for original text, shuffling in paragraph level, and word level for *"On The Origin of species"*. According to the figure, in large scales, scaling exponent of the original text is 0.73, for the case of shuffled paragraphs $\alpha = 0.63$, and for shuffling in word level $\alpha = 0.5$. For the graphic presentation we depict only some of the levels in the Fig. 8 and inscribe the scaling exponent for shuffled texts of other levels in Table 3. For another instance Fig. 8b shows the result of DFA1 for the original text, shuffling in paragraph level, and word level for *"Relativity"*. In large scales, scaling exponent of the original text is 0.80, for shuffling in paragraph level $\alpha = 0.62$,

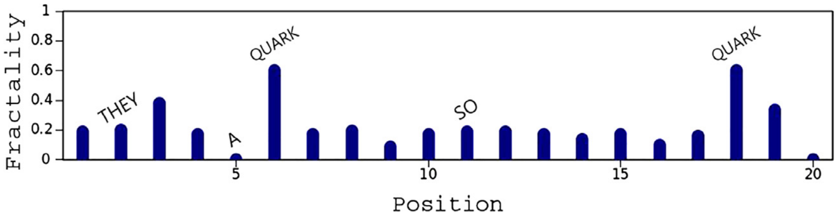


Fig. 6. A schematic of constructing fractality time series from a sample text chosen from *The First Three Minutes*. The sample text is *"IF THEY CONSIST OF A QUARK AND AN ANTIQUARK AND SO ON BUT DESPITE THIS SUCCESS THE QUARK MODEL PRESENTS"*.

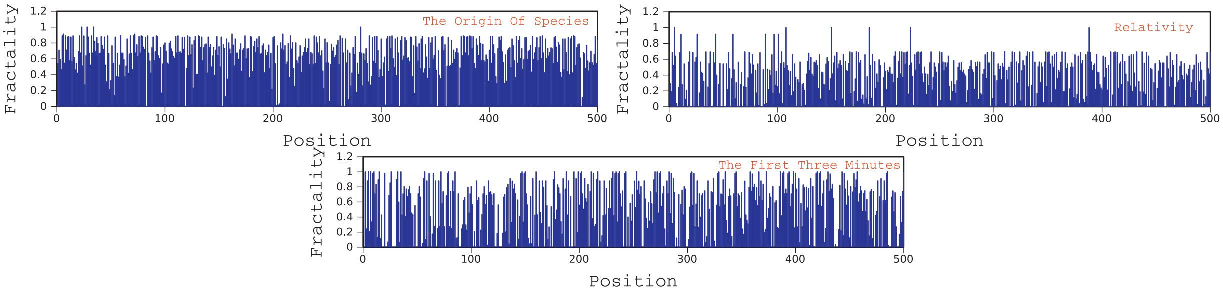
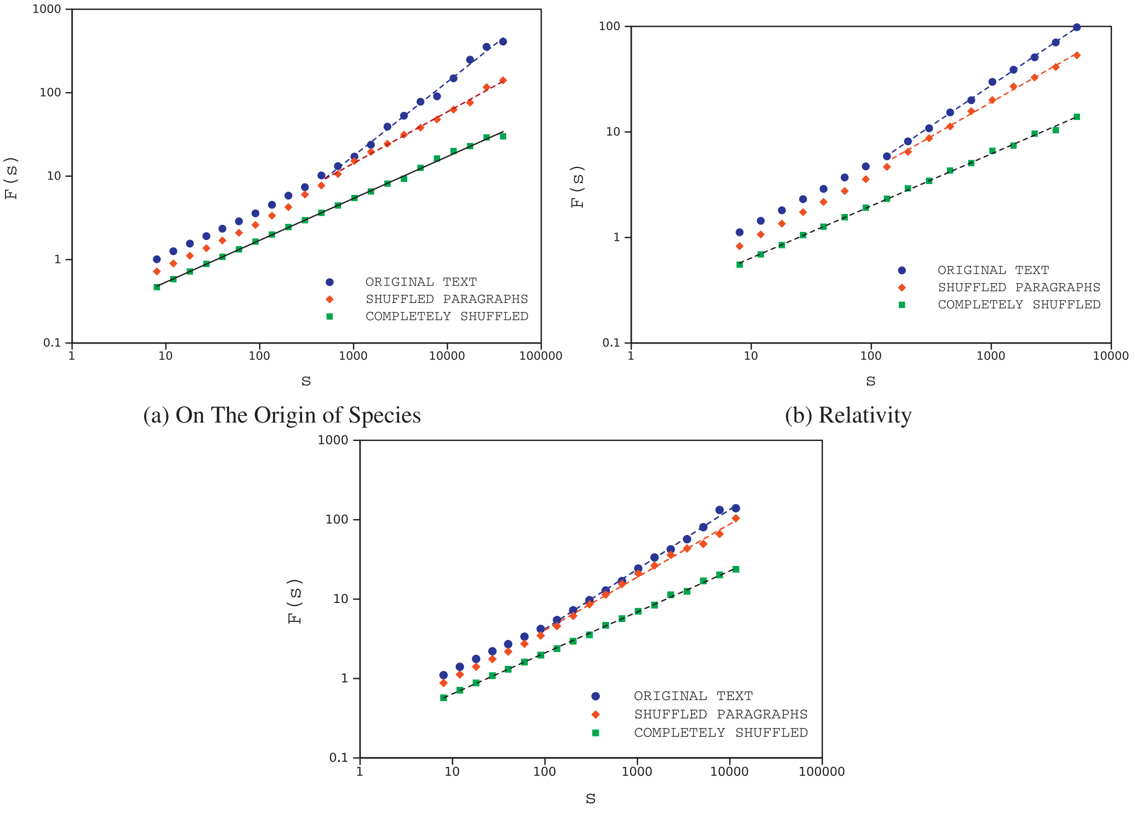


Fig. 7. Importance time series for three books. Normalized *Fractality* time series are shown for the first 500 words of, *On The Origin of Species*, *The First Three Minutes*, and *Relativity* correspondingly.



(a) On The Origin of Species

(b) Relativity

(c) The First Three Minutes

Fig. 8. DFA1 for Fractality time series of our three sample texts. By gradually shuffling the text, α exponent of fractality time series decreases slowly. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

and for shuffling in word level $\alpha = 0.5$ again. According to result of DFA1 for "The First Three Minutes" in Fig. 8c, α exponent for the original text is 0.72, for shuffling in paragraph level is 0.71 and for shuffling in word level is $= 0.5$ again. Table 3 shows scaling exponents for our sample books in different levels of shuffling, using DFA method with different orders of detrending: DFA1, DFA2, DFA3, DFA4, and DFA5. The scaling exponents by using DMA method is also included in the table. Table 3 demonstrates the most important outcome of the procedure we applied to the sample texts. This table represents that the method is independent of the properties of texts and also is independent of fitting protocols. As is seen, for all of detrended methods shuffling a text level by level results in step-by-step decrease in the value of scaling exponent. The text with shuffled chapters for the book *The First Three Minutes* is an exception. The chapters of this text is written independently. So, the shuffling in chapters level do not decrease the coherence of the text. This point is also seen in the estimated coherence by IBM model (Table 5).

Table 3
The scaling exponent of *Fractality* time series of our three sample books, using DFA1, DFA2, DFA3, DFA4, DFA5, and DMA methods. α exponent decreases step by step with shuffling of texts level by level. The only exception occurs for The *The First Three Minutes* with shuffled chapters. The scaling exponent in this level is larger than the case of the original text for all orders of DFA and also for DMA.

Book's name	State	α					
		DFA1	DFA2	DFA3	DFA4	DFA5	DMA
The Origin of Species	Original Text	0.73	0.74	0.72	0.71	0.70	0.81
	Text with Shuffled Chapters	0.71	0.72	0.71	0.70	0.69	0.80
	Text with shuffled Paragraphs	0.63	0.61	0.63	0.63	0.63	0.55
	Text with shuffled Sentences	0.51	0.51	0.52	0.53	0.53	0.50
	Shuffled Text	0.50	0.50	0.50	0.49	0.49	0.47
Relativity	Original Text	0.80	0.80	0.80	0.77	0.74	0.80
	Text with Shuffled Chapters	0.75	0.78	0.73	0.73	0.73	0.77
	Text with shuffled Paragraphs	0.62	0.65	0.65	0.66	0.66	0.66
	Text with shuffled Sentences	0.62	0.65	0.63	0.64	0.65	0.63
	Shuffled Text	0.50	0.50	0.50	0.50	0.50	0.50
The First Three Minutes	Original Text	0.72	0.72	0.71	0.70	0.70	0.75
	Text with Shuffled Chapters	0.74	0.74	0.75	0.74	0.72	0.76
	Text with shuffled Paragraphs	0.71	0.71	0.70	0.72	0.72	0.70
	Text with shuffled Sentences	0.55	0.54	0.55	0.56	0.57	0.53
	Shuffled Text	0.50	0.51	0.50	0.51	0.51	0.48

Table 4
The entity grid scores for our sample books by using three models. *EG* stands for generative entity grid model [11], *EGF* is the entity grid with features [34], and IBM stands for the IBM model learned on Wall Street Journal data [35]. All the numbers in the table should be multiplied in 10^4.

Book's name	State	EG	EGF	IBM
On the Origin of Species	Original Text	−181.048	−167.425	−53.679
	Text with Shuffled Chapters	−182.633	−168.933	−53.6968
	Text with shuffled Paragraphs	−182.551	−168.891	−53.8277
	Text with shuffled Sentences	−183.655	−169.444	−55.0731
	Shuffled Text	−240.898	−221.169	−67.3759
Relativity	Original Text	−9.81868	−9.24884	−9.42337
	Text with Shuffled Chapters	−9.86012	−9.88242	−9.41931
	Text with shuffled Paragraphs	−9.89874	−9.32464	−9.47683
	Text with shuffled Sentences	−10.0243	−9.45495	−9.7719
	Shuffled Text	−13.0876	−12.0511	−11.3937
The First Three Minutes	Original Text	−22.7213	−21.2112	−15.6273
	Text with Shuffled Chapters	−22.8361	−21.3217	−15.6156
	Text with shuffled Paragraphs	−22.8808	−21.3624	−15.6601
	Text with shuffled Sentences	−22.9879	−21.4845	−16.1542
	Shuffled Text	−30.4243	−27.844	−18.6402

The reason for this gradual decrease in α exponent is that a text is not just a bunch of words, but a collection of well- positioned words. So, the meaning of a text is not only because of words it contains, but also because of the positions of words along the text. If we shuffle some parts of a text, we certainly lose a piece of the meaning and the α exponent decreases as well. If we continue to shuffle the text parts in lower levels, more of the meaning will be lost and α exponent will decrease more. Finally if we continue to shuffle all of the text, the meaning will be completely lost and α exponent will be equal to 0.5.

As we mentioned before, with shuffling a text, the coherence of the text decreases. On the other hand, we found that shuffling process decreases scaling exponent of a text. As a result, scaling exponent is related to coherence of texts. A text with higher α exponent is more coherent. So, scaling exponent of a text time series (STT) could be an coherence indicator. With the help of this method we can quantify coherence of texts and quantitatively compare coherence of different texts from an author or texts with the same subject from different authors.

Table 5
The Scaling exponent of *Fractality, C value, Entropy* and *Frequency* time series for our three sample books, using DFA1 method. α exponent decreases little by little with gradual shuffling of texts.

Book's name	State	α			
		Fractality	*Cvalue*	Entropy	Frequency
The Origin of Species	Original Text	0.73	0.84	0.67	0.65
	Text with Shuffled Chapters	0.71	0.83	0.64	0.61
	Text with shuffled Paragraphs	0.63	0.59	0.53	0.56
	Text with shuffled Sentences	0.51	0.54	0.50	0.52
	Shuffled Text	0.50	0.49	0.49	0.50
Relativity	Original Text	0.80	0.80	0.66	0.66
	Text with Shuffled Chapters	0.75	0.78	0.66	0.66
	Text with shuffled Paragraphs	0.62	0.74	0.54	0.58
	Text with shuffled Sentences	0.62	0.65	0.51	0.49
	Shuffled Text	0.50	0.51	0.50	0.49
The First Three Minutes	Original Text	0.72	0.73	0.70	0.64
	Text with Shuffled Chapters	0.74	0.74	0.71	0.66
	Text with shuffled Paragraphs	0.71	0.67	0.60	0.54
	Text with shuffled Sentences	0.55	0.53	0.51	0.53
	Shuffled Text	0.50	0.52	0.49	0.50

To examine the accuracy of STT method, we repeated the procedure with other importance indexes. In addition to fractality measure, there are other methods that are used in word ranking tasks; including C value, Entropy and Frequency [25–27]. We introduce these importance measures in Appendix section. After assigning importance values to word-types by one of these methods, we construct C value time series, Entropy time series and Frequency time series from our sample texts and the various shuffled versions. It was again seen that the scaling exponent of text importance time series decrease with gradual shuffling of the time series. We inscribe scaling exponent for importance time series using C value, Entropy and Frequency measures in Table 5 in the Appendix section.

By usage of correlation coefficient of text time series we have been able to calculate the global coherence of a text automatically.

This method is ab initio and it is not even needed to know the subject of the text in advance.

3.1. Comparison with other methods

Unfortunately there is no precise model to evaluate our method with, but there are some models which are powerful enough in task of coherence calculation. So, to give an intuition about the result of our STT method, We compare the results with some entity grid based coherence models. Before calculating coherence scores with these models, some pre-processing should be done to the sample texts. First the sentences should be split separately (one sentence in each line) and secondly the sample texts should be parsed by a parser. We benefit of Stanford Parser [33] for this purpose. We select three coherence models to compare our method with: the generative entity grid model (EG) [11], entity grid with features (EGF) [34] and the IBM model [35]. In generative entity grid model, entity columns are considered independently and each one is modeled with a Markov chain. The second model is an extension of the entity grid which add number of entity-specific features like discourse prominence, named entity type and coreference features to distinguish between important and unimportant entities. And the IBM model which is learned on Wall Street Journal data assumes that existence of certain words in a sentence, cause the tendency of the usage of certain words in the translation of that sentence in other languages. All of these coherence models are freely available from Brown Coherence Toolkit [36]. The coherence scores which is calculated by these three models are brought in Table 4.

To compare these models with our STT method, we use Spearman's rank correlation coefficient or Spearman's r_s [37]:

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)}. \tag{9}$$

Where r_s is the rank correlation coefficient between two sets of variables X and Y, n is the total number of each variable set and $d_i = X_i - Y_i$ is the difference between the two ranks of variable sets. This coefficient is used to measure the strength of the rank correlation between two variables. The maximum value of this quantity is 1. The coefficient equals to 1 occurs when each of the two variables indicates exactly the same rank as the other. Sample texts and their shuffled versions are ranked due to their coherence values by four coherence models (EG, EGF, IBM and STT). To calculate the rank correlation coefficient, we use all three books and their shuffled version to derive the r_s results.

The rank correlation coefficient associate with each two models is brought in the matrix below. EG refers to the generative entity grid model, EGF is the entity grid model with features and STT refers to scaling exponent of text time series method.

$$r_s = \begin{array}{c} \\ EG \\ EGF \\ IBM \\ STT \end{array} \begin{array}{c} \begin{array}{cccc} EG & EGF & IBM & STT \end{array} \\ \begin{pmatrix} 1.000 & 0.989 & 0.989 & 0.993 \\ 0.989 & 1.000 & 0.971 & 0.982 \\ 0.989 & 0.971 & 1.000 & 0.996 \\ 0.993 & 0.982 & 0.996 & 1.000 \end{pmatrix} \end{array}$$

As is seen from this matrix, all the methods operate almost similarly in document ranking task and are strongly rank correlated with each other. In compare with other methods, the STT needs no pre-processing and do not need to use a parser. We just apply the method on raw texts. Moreover our method is faster; the time of our procedure is about some seconds up to one minute. But the other methods need some minutes up to hours. Specially the pre-processing tasks such as parsing takes so much times in order of some hours. Also some parsers have some restrictions about sentence size. The parser we used could not parse the sentences which contains more than 70 words. The advantage of STT method is its fast procedure and more importantly its lack of necessity to any pre-processing task.

4. Conclusion

In this paper we suggest a novel method for quantifying global coherence of texts. The method consists of three steps: 1- An importance value is assigned to each word-type of texts by using one of the importance measures like fractality, C value and Entropy. 2- Importance time series is constructed from texts by replacing each word with its importance value. 3- By DFA method, importance time series of texts is analyzed and long-range correlations is detected. In studying the correlations in text time series, we observed an interesting point; The importance time series of texts have two different scaling exponents, one for small scales and another for large scales. We believe that this behavior stands in trace of universality and call it bi-scale behavior of texts; It is found that fractality time series of texts are short-range correlated in small scales but, they exhibit long-range correlation in large scales. By analyzing importance time series of a text we can calculate the scaling exponent of the text. The scaling exponent of text time series is the indicator of text's global coherence. Texts with higher scaling exponents are more coherent. The advantage of this method is that it is ab initio (it does not even need a corpus, stemming or removing stop words) and also, we do not need to know the subject of a text beforehand. Moreover, the calculations do not need any preprocessing and even parsing and computational efforts. The computations are straightforward and almost easy to follow and also time-saving.

For our future research, we will use the STT method for evaluation of text summarization and question answering tasks. We will also use the fractality time series to measure semantic information content of text.

Acknowledgment

We would like to show our gratitude to Dr. Micha Elsner (Department of Linguistics, Ohio State University), for preparing coherence scores of entity grid based coherence models.

Appendix A. Calculating coherence by using other importance measures

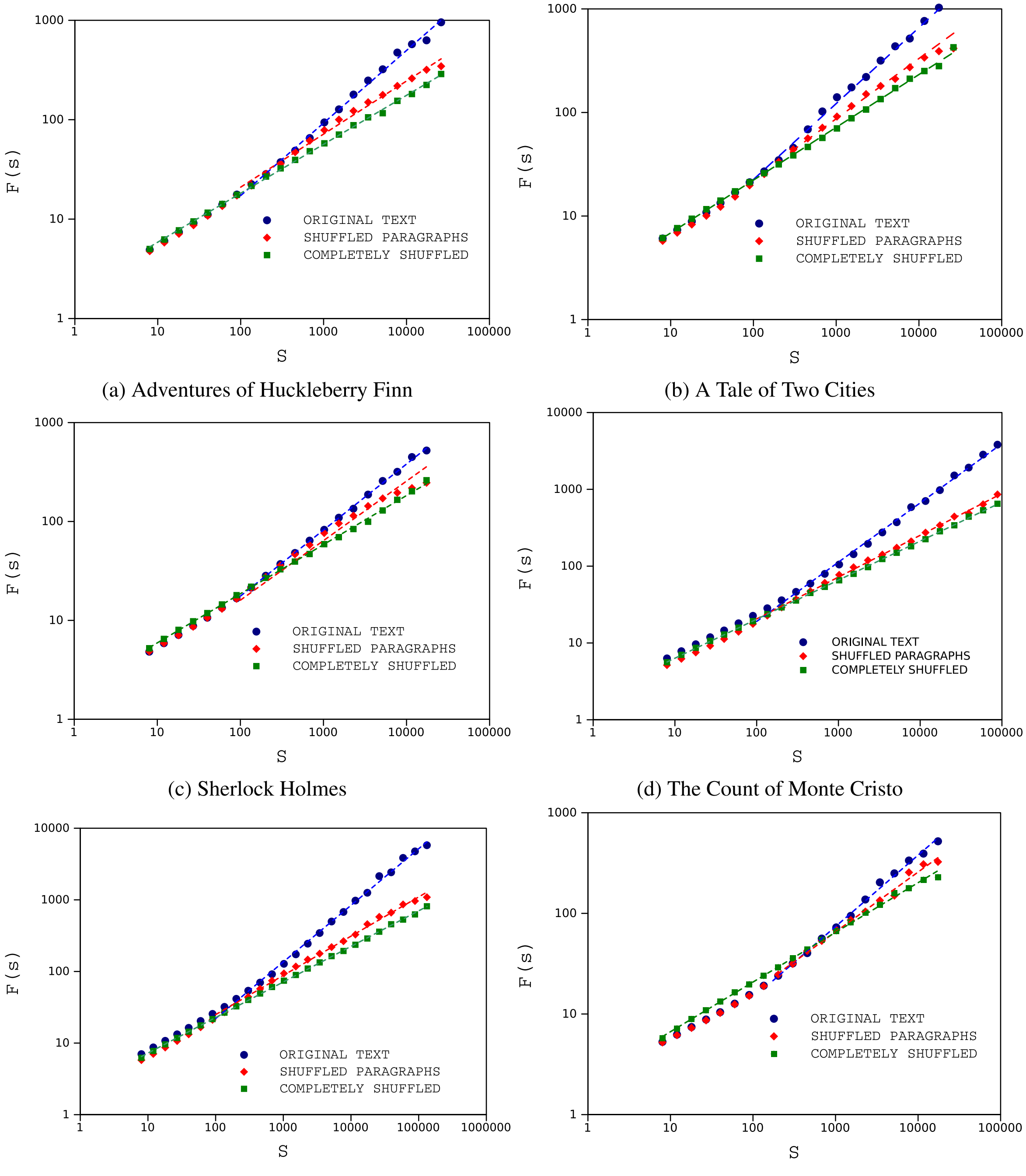
In addition to the fractality method, there are some other methods which assign the importance values to words in a text. *Frequency* is a raw method which uses the frequency of a word (i.e. the number of times a word is repeated in a text) as a measure of its importance. *C Value* [26] and *Entropy* [27] are two other methods with high efficiency in word ranking [6,7]. Here we only discuss how these methods calculate importance values and refer the interested reader to the original papers for complete technical details [26,27].

C Value method uses the standard deviation of distances between two consecutive occurrences of a word as an index for its importance [26].

$$C(w) = \sqrt{\nu}(1 + 2.8\nu^{-0.865})\left\{\frac{\sigma(w)}{\sqrt{1-\nu/N}} - \frac{2\nu-1}{2\nu+1}\right\}, \tag{A.1}$$

where ν is frequency of the word type w, N is the total number of words in the text and $\sigma(w)$ shows the normalized standard deviation for inter-word distances.

In *Entropy* method, text is divided into some parts; namely the chapters, using the Shannon definition of entropy we can deter-



(a) Adventures of Huckleberry Finn

(b) A Tale of Two Cities

(c) Sherlock Holmes

(d) The Count of Monte Cristo

(e) War and Peace

(f) Frankenstein

Fig. 9. DFA1 for Fractality time series of some sample texts. By gradually shuffling the text, α exponent of fractality time series decreases slowly.

mine the importance of any word [27],

$$E(w) = \frac{2\nu \ln P}{p-1}\left(1 + \frac{1}{\ln P}\sum_{i=1}^{P} p_i(w)\ln p_i(w)\right), \tag{A.2}$$

where ν again shows frequency of the word w, P is the number of text segments and $p_i(w)$ is the probability of occurrence of a word type in the ith part. p_i is equal to ν_i/ν where ν_i enumerates the frequency of w in the part i.

After using these methods to assign importance values to words, C value, Entropy, and Frequency time series can be constructed by replacing each word with its importance value. Table 5 shows scaling exponents of *Fractality*, *C value*, *Entropy*, and *Frequency* time series of *On the Origin of Species* in different levels of shuffling, using DFA1 method. As is seen in the table, shuffling a text step by step results in gradual decrease of scaling exponent of its importance time series.

According to the figure, scaling exponent of the sample book decreases with gradual shuffling of the time series. So, there is no difference in choice of importance indexes and the global coherence of a text could be quantified by analyzing importance time series of the text.

Appendix B

To extend the universe of our work we study many texts from fiction and non-fiction genres. In Fig. 9 we plot the diagram of $F(s)$ versus s for some of these texts. As is seen from the figure by shuffling a text step by step the correlation exponent decreases.

References

[1] M. De Gruyter, Cognitive Linguistics in Action: From Theory to Application and Back, Walter de Gruyter GmbH and Co. KG, Berlin, New York, 2010.
[2] M. Louwerse, A.C. Graesser, Coherence in discourse, in: P. Strazny (Ed.), Encyclopedia of Linguistics, Fitzroy Dearborn, Chicago, 2005, pp. 216–218.
[3] H. Whitaker, Concise Encyclopedia of Brain and Language, Elsevier, Oxford, 2010.
[4] E. Fernndez, H.S. Cairns, Fundamentals of Psycholinguistics, John Wiley and Sons Ltd, United Kingdom, 2011.
[5] D. Geeraerts, Theories of Lexical Semantics, Oxford University Press, Oxford, 2010.
[6] E. Najafi, A.H. Darooneh, The fractal patterns of words in a text: a method for automatic keyword extraction, PLoS ONE 10 (6) (2015) E0130617, doi:10.1371/journal.pone.0130617.
[7] A. Mehri, A.H. Darooneh, The role of entropy in word ranking, Phys A 390 (2011) 3157–3163.
[8] C. Peng, S. Halvin, H. Stanley, G. L., Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series, Chaos 5 (1) (1995) 7–82.
[9] J.W. Kantelhardt, Fractal and multifractal time series, in: Mathematics of Complexity and Dynamical Systems, Springer, 2011, pp. 463–487.
[10] Y. Shao, G. Gu, Z. Jiang, W. Zhou, D. Sornette, Comparing the performance of FA, DFA and DMA using different synthetic long-range correlated time series, Sci. Rep. 2 (2012) 835.
[11] R. Barzilay, M. Lapata, Modeling local coherence: an entity-based approach, Comput. Linguist. 34 (2008) 1–34.
[12] K. Landauer, D. McNamara, S. Dennis, W. Kintsch, Handbook of Latent Semantic Analysis, Psychology Press, 2013. ISBN : 1135603278, 9781135603274.
[13] M. Dias, V. Feltrim, T.A.A. Padro, Using rhetorical structure theory and entity grids to automatically evaluate local coherence in texts, in: Computational Processing of the Portuguese Language, Lecture Notes in Computer Science, vol. 8775, 2014, pp. 232–243.
[14] Z. Lin, H. Ng, M.Y. Kan, Automatically evaluating text coherence using discourse relations, in: Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA, vol. 1, 2011, pp. 997–1006.
[15] M.A. Halliday, R. Hasan, Cohesion in english, English Language Series, Longman Pub Group, 1976.
[16] B. Bamberg, Coherence and cohesion: what are they and how are they achieved? Coll. Compos. Commun. 34 (4) (1983) 417–429.
[17] T. Dijk, W. Kintsch, Strategics in Discourse Comprehension, Academic Press, New York, 1983.
[18] P. Foltz, W. Kintsch, T.K. Landauer, The measurement of textual coherence with latent semantic analysis, Discourse Processes 25 (1998) 285–307.
[19] P. Kanerva, J. Kristoferson, A. Holst, Random indexing of text samples for latent semantic analysis, in: Proceeding of 22nd Annual Conference of the Cognitive Science Society, 2000.
[20] K. Filippova, M. Strube, Extending the entity-grid coherence model to semantically related entities, in: Proceeding of the Eleventh European Workshop on Natural Language Generations, 2007, pp. 139–142.
[21] J. Burstein, J. Tetreault, S. Andreyev, Using entity-based features to model coherence in student essays, in: Proceeding of the 2010 Annual Conference of thec North American Chapter of the Association for Computational Linguistics, Human Language Technologies, 2010, pp. 681–684.
[22] R. Iida, T. Tokunaga, A metric for evaluating discourse coherence based on coreference resolution, in: Proceedings of the COLING 2012: Posters, Mumbai, India, 2012, pp. 483–494.
[23] J. Gouyet, Physics and Fractal Structures, Masson Springer, New york, 1996.
[24] D. Higgins, J. Burstein, D. Marcu, C. Gentile, Evaluating multiple aspects of coherence in student essays, in: Proceeding of the NAACL, 2004, pp. 185–192.
[25] H.P. Luhn, The automatic creation of literature abstracts, IBM J. Res. Dev. 2 (1958) 159–165.
[26] P. Carpena, P. Bernaola-Galvan, M. Hackenberg, A. Coronado, J.L. Oliver, Level statistics of words: finding keywords in literary texts and symbolic sequences, Phys. Rev. E 79 (2009) 035102.
[27] J. Herrera, P.A. Pury, Statistical keyword detection in literary corpora, Eur. Phys. J. B 63 (2008) 135.
[28] S. Arianos, A. Carbone, Detrending moving average algorithm: a closed-form approximation of the scaling law, Phys. A 382 (2007) 9–15.
[29] A. Schumann, J.W. Kantelhardt, Multifractal moving average analysis and test of multifractal model with tuned correlations, Phys. A 390 (2011) 2637–2654.
[30] C. Darwin, On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life, in: Nature, John Murray, London, 1859, p. 5.
[31] A. Einstein, Relativity: The Special and General Theory, H. Holt and Company, New York, 1920.
[32] S. Weinberg, The first three minutes: a modern view of the origin of the universe, Basic Books, 1976.
[33] D. Klein, C.D. Manning, Accurate unlexicalized parsing, in: Proceedings of the 41st Meeting of the Association for Computational Linguistics, 2003, pp. 423–430.
[34] M. Elsner, E. Charniak, Extending the entity grid with entity-specific features, in: Proceeding of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, Portland, Oregon, 2011, pp. 125–129.
[35] R. Soricut, D. Marcu, Discourse generation using utility-trained coherence models, in: Proceeding of the Association for Computational Linguistics Conference (ACL-2006), 2006.
[36] https://bitbucket.org/melsner/browncoherence.
[37] http://mathworld.wolfram.com/SpearmanRankCorrelationCoefficient.htm.